THE UNIVERSITY
OF MICHIGAN

SEP 12 1962

ENGINEERING
LIBRARY

АВТОМАТИКА
И
ТЕЛЕМЕХАНИКА

Vol. 22, No. 6, June, 1961

Translation Published December, 1961

SOVIET INSTRUMENTATION AND
CONTROL TRANSLATION SERIES

Automation and Remote Control

(The Soviet Journal *Avtomatika i Telemekhanika* in English Translation)

■ This translation of a Soviet journal on automatic control is published as a service to American science and industry. It is sponsored by the Instrument Society of America under a grant in aid from the National Science Foundation, continuing a program initiated by the Massachusetts Institute of Technology.

SOVIET INSTRUMENTATION AND CONTROL TRANSLATION SERIES

Instrument Society of America

Philip A. Sprague
President
Dr. Ralph H. Tripp
Past President
Nathan Cohn
President-elect-Secretary
Henry J. Noebels
Dept. Vice President
Dr. Benjamin W. Thomas
Dept. Vice President-elect
E. A. Adler
Dept. Vice President
Francis S. Hoag
Dept. Vice President
John J. McDonald
Dept. Vice President
John R. Mahoney
Dept. Vice President-elect
John C. Koch
Treasurer
Alonzo B. Parsons
Dist. I Vice President
H. Kirk Fallin
Dist. II Vice President
Harold J. Bowman
Dist. III Vice President
Fred R. Gilmer
Dist. IV Vice President
Charles H. Callier
Dist. V Vice President
Otto J. Lessa
Dist. VI Vice President
J. Howard Park, III
Dist. VII Vice President
C. Roy Horton
Dist. VIII Vice President
Max Curtis
Dist. IX Vice President
Kenneth S. Vriesen
Dist. X Vice President
Allan E. Lee
Dist. XI Vice President

International Headquarters

William H. Kushnick
Executive Director
Charles W. Covey
Editor, ISA Journal
Herbert S. Kindler
Director, Tech. & Educ. Services

ISA Publications Committee

Charles O. Badgett, *Chairman*
Jere E. Brophy, George A. Larsen, Joshua Stern
Dr. Enoch J. Durbin, Thomas G. MacAnespie, Frank S. Swaney
Prof. Richard W. Jones, John E. Read, Richard A. Terry

Translations Advisory Board of the Publications Committee

Jere E. Brophy, *Chairman*
T. J. Higgins, S. G. Eskin, G. Werbizky

■ This translation of the Soviet Journal *Avtomatika i Telemekhanika* is published and distributed at nominal subscription rates under a grant in aid to the Instrument Society of America from the National Science Foundation. This translated journal, and others in the Series (see back cover), will enable American scientists and engineers to be informed of work in the fields of instrumentation, measurement techniques, and automatic control reported in the Soviet Union.

The original Russian articles are translated by competent technical personnel. The translations are on a cover-to-cover basis and the Instrument Society of America and its translators propose to translate faithfully all of the scientific material in *Avtomatika i Telemekhanika*, permitting readers to appraise for themselves the scope, status, and importance of the Soviet work. All views expressed in the translated material are intended to be those of the original authors and not those of the translators nor the Instrument Society of America.

Publication of *Avtomatika i Telemekhanika* in English translation started under the present auspices in April, 1958, with Russian Vol. 18, No. 1 of January, 1957. The program has been continued with the translation and printing of the 1958-1961 issues.

Transliteration of the names of Russian authors follows the system known as the British Standard. This system has recently achieved wide adoption in the United Kingdom, and is currently being adopted by a large number of scientific journals in the United States.

Readers are invited to submit to the Instrument Society of America comments on the quality of the translations and the content of the articles. Pertinent correspondence will be published in the Society's monthly publication, the ISA JOURNAL. Space will also be made available in the ISA JOURNAL for such replies as may be received from Russian authors to comments or questions by the readers.

1961 Volume 22 Subscription Prices:

Per year (12 issues), starting with Vol. 22, No. 1

General:	*United States and Canada*	$35.00
	Elsewhere	38.00
Libraries of nonprofit academic institutions:		
	United States and Canada	$17.50
	Elsewhere	20.50
Single issues to everyone, each		$ 6.00

1957 Volume 18, 1958 Volume 19, 1959 Volume 20, and 1960 Volume 21 issues also available. Prices upon request.

See back cover for combined subscription to entire Series.

Subscriptions and requests for information on back issues should be addressed to the:

Instrument Society of America
530 William Penn Place, Pittsburgh 19, Penna.

Translated and printed by Consultants Bureau Enterprises, Inc.

Copyright © 1961 by the Instrument Society of America

Automation and Remote Control

A translation of Avtomatika i Telemekhanika, a publication of the Academy of Sciences of the USSR

EDITORIAL BOARD OF AVTOMATIKA I TELEMEKHANIKA

D. I. Ageikin
M. A. Aizerman
A. B. Chelyustkin *(Assoc. Editor)*
E. G. Dudnikov
N. Ya. Festa
V. A. Ilin
A. G. Iosuf'yan
V. V. Karibskii
A. V. Khramoi
B. Ya. Kogan
V. S. Kulebakin
S. A. Lebedev
A. Ya. Lerner
A. M. Letov *(Assoc. Editor)*
V. S. Malov
B. N. Petrov
Yu. P. Portnov-Sokolov
B. S. Sotskov
A. A. Tal' *(Corresp. Secretary)*
V. A. Trapeznikov *(Editor in Chief)*
Ya. Z. Tsypkin
G. M. Ulanov
A. A. Voronov
S. V. Yablonskii

Vol. 22, No. 6 Russian Original Dated June, 1961 December, 1961

CONTENTS

	PAGE	RUSS. PAGE
On the Realization of Motion Along a Given Trajectory. E. A. Barbashin	587	681
The Analytical Design of Control Systems. Ya. Kurtsveil'	593	688
The Extrapolation-Gradient Method of Searching for the Minimum of a Quadratic Function. R. É. Vinograd and Yu. V. Geronimus	599	696
Investigation of Stability of Periodic States in Nonlinear Pulse Automatic Systems. Ya. Z. Tsypkin	614	711
The Statistical Theory of Amplitude Quantization. A. A. Kosyakin	624	722
Extremal Reception of Signals. A. A. Krasovskii	631	730
A Method of Plotting Amplitude and Phase Frequency Characteristics with the Help of a Generalized Family of Logarithmic Characteristics. D. E. Polonnikov	639	739
The Algorithmic Unsolvability of the Problem of Recognizing the Representability of Recursive Events in Finite Automata. M. A. Aizerman, L. A. Gusev, L. I. Rozonoér, I. M. Smirnova, and A. A. Tal'	646	748
Canonical Method of Switching Circuit Synthesis. A. Sh. Blokh	652	756
Relay-Type Statistical Automata and Certain Methods of Their Investigation. I. I. Perel'man	661	765
Time Quantization Error in Automatic Control. S. M. Mandel'shtam	675	780
Theory of a Two-Core Magnetic Amplifier with Rectification of the Even Harmonics by Means of a Nonlinear Resistance. Dai Tse-Hsin	681	787
The Distribution of the Peak Durations for Pulse Noise at the Output of a Remote Control Unit. L. B. Venchkovskii	689	795
Some Cases of Similarity of Transient Processes in Single-Loop Nonlinear Control Systems. V. T. Dedesh	695	801
On the Utilization of Digital Computers for Differentiating and Smoothing Out Sequences Containing Random Disturbances. A. N. Pokrovskii	697	803
On Transient Response Conditions in a Magnetic Amplifier with Feedback and an Inductive Load that is Connected through a Rectifier. V. A. Sokolov	701	807
CHRONICLE		
All-Union Conference on Contactless Magnetic Elements which Are Used in Automation and Computer Techniques. G. V. Subbotina	708	811

SOVIET INSTRUMENTATION AND
CONTROL TRANSLATION SERIES

Automation and Remote Control

A translation of Avtomatika i Telemekhanika, a publication of the Academy of Sciences of the USSR

EDITORIAL BOARD OF AVTOMATIKA I TELEMEKHANIKA

Vol. 22, No. 6 December, 1961

CONTENTS

PAGE PAGE

ON THE REALIZATION OF MOTION ALONG A GIVEN TRAJECTORY

E. A. Barbashin

(Sverdlovsk)

Translated from Avtomatika i Telemekhanika, Vol. 27, No. 6,
pp. 681-687, June, 1961
Original article submitted December 12, 1960

Methods are given of selecting system parameters which enable the approximate realization of a given process. Conditions are derived for which the problem can be reduced to a problem of the theory of RMS approximations.

Let us examine the system of differential equations

$$\frac{dx_i}{dt} = f_i(x_1, \ldots, x_n, t) + \varphi_i(c_1, \ldots, c_k, y_1, \ldots, y_m, t) \quad (i = 1, 2, \ldots, n), \tag{1}$$

where $c_1, \ldots, c_k$ are the constant parameters, and $y_1, \ldots, y_m$ are functions of time t.

Together with the system (1) we assume over the interval $0 \leq t \leq T$, a trajectory $x_i = \psi_i(t)$ $(i = 1, 2, \ldots, n)$ which, in general, does not satisfy the system (1). Our task will be to find the parameters $c_1, \ldots, c_k$ or the functions $y_1(t), \ldots, y_m(t)$, such that the solution $x_i(t)$ of system (1), defined by the initial conditions $x_i(0) = x_i^0$ (the differences $x_i^0 - \psi_i(0)$ are sufficiently small) be an approximate realization of the motion along the given trajectory $x_i = \psi_i(t)$ at $0 \leq t \leq T$. As will be shown in the following, it is advantageous to take as a measure of approximation the root-mean-square value of the error at which the given trajectory satisfies the system (1). Thus, we shall select the control parameters $c_1, \ldots, c_k$ (task A) or the control functions $y_1(t), \ldots, y_m(t)$ (task B), or, finally, both the control parameters and control functions (task C), such as to have a minimum RMS error involved. Problems of the aforementioned type for certain particular cases have been solved in the works.

The present work substantiates the proposed procedure for the nonlinear system (1), while in the case where the functions $\varphi_i(c_1, \ldots, c_k, y_1, \ldots, y_m, t)$ are linear functions of the parameters $c_1, \ldots, c_k$ and functions $y_1, \ldots, y_m$, it presents efficient techniques for the selection of optimum parameter values.

1. Substantiation of the Method

Substituting in system (1) the variables $z_i = x_i - \psi_i(t)$ $(i = 1, 2, \ldots, n)$, we get the system

$$\frac{dz_i}{dt} = f_i[z_1 + \psi_1(t), \ldots, z_n + \psi_n(t), t] + \varphi_i(c_1, \ldots, c_k, y_1, \ldots, y_m, t) - \psi_i'(t).$$

Introducing the denotations

$$r_i(t) = \psi_i'(t) - f_i[\psi_1(t), \ldots, \psi_n(t), t], \quad Z_i(z_1, \ldots, z_n, t) =$$

$$= f_i[z_1 + \psi_1(t), \ldots, z_n + \psi_n(t), t] - f_i[\psi_1(t), \ldots, \psi_n(t), t] \quad (i = 1, 2, \ldots, n),$$

we write the new system in the form

$$\frac{dz_i}{dt} = Z_i(z_1, \ldots, z_n, t) + \varphi_i(c_1, \ldots, c_k, y_1, \ldots, y_m, t) - r_i(t), \tag{2}$$

where it is readily conceived that $Z_i(0, \ldots, 0, t) = 0$ $(i = 1, 2, \ldots, n)$.

The problem then reduces to the approximate realization of motion along the trajectory $z_i = 0$ $(i = 1, 2, \ldots, n)$ at $0 \leq t \leq T$.

Let us assume that the functions $Z_i(z_1, \ldots, z_n, t)$ are continuous for all arguments, and that they satisfy the Lipschitz conditions

$$|Z_i(z_1, \ldots, z_n, t) - Z_i(z_1^0, \ldots, z_n^0, t)| \leqslant L \sum_{k=1}^{n} |z_k - z_k^0|. \tag{3}$$

We shall not impose upon the function $\varphi_i(c_1, \ldots, c_k, y_1, \ldots, y_m, t)$ any specific limitations, except those which are necessary for the existence and uniqueness of the solutions of system (2) in the region D.

Following the course initially set, we select $c_1, \ldots, c_k, y_1(t) \ldots, y_m(t)$, such as to have the line $z_i = 0$ $(i = 1, 2, \ldots, n)$ satisfy the system (2) with a minimum RMS error. This means that the selection of the control parameters or control functions should be such as to provide for an approximation of the function $r_i(t)$ by the functions $\varphi_1(c_1, \ldots, c_k, y_1, \ldots, y_m, t)$ at a minimum RMS error in the interval $0 \leq t \leq T$.

Let us note that with the transfer from system (1) to system (2), the magnitude of the deviation of the actual from the given trajectory is determined by the deviation from zero of the solution $z_i(t)$ of system (2), defined by the initial conditions $z_i(0) = z_i^0$, where the numbers z_i^0 are sufficiently close to zero.

Let us introduce the denotations:

$$\Delta^2(t) = \sum_{i=1}^{n} [\varphi_i(c_1, \ldots, c_k, y_1, \ldots, y_m, t) - r_i(t)]^2, \quad H^2 = \int_0^T \Delta^2(t)\, dt. \tag{4}$$

Theorem 1. Let $z_i(t)$ be the solution of system (2), defined by the conditions $z_i(0) = z_i^0; |z_i^0| < \delta$ $(i = 1, 2, \ldots, n)$. If, at $0 \leq t \leq T$, this solution does not surpass the limits of the D region, then there takes place the estimate

$$|z_i(t)| \leqslant \delta e^{nLT} + \left(\frac{e^{2nLT} - 1}{2nL}\right)^{1/2} H \qquad (i = 1, 2, \ldots, n). \tag{5}$$

The estimate (5) shows that for a sufficiently small H, the trajectory in the interval $0 \leq t \leq T$ can be approximated with any degree of accuracy in the sense that the maximum of the deviation of the actual from the desired trajectory can be made arbitrarily small. However, if T is sufficiently large, the deviation of the actual from the given trajectory may become substantial, since the growth of the maximum in the general case may proceed according to the exponential law. Hence, it is important to define the conditions which provide for the absence of such an undesirable growth in error. Let us formulate the said conditions.

Let

$$h^2 = \sup_{0 \leqslant t \leqslant T-1} \int_t^{t+1} \Delta^2(t)\, dt, \tag{6}$$

where $\Delta(t)$ is determined by the equality (4).

Together with the system (2) we examine the system

$$\frac{dx_i}{dt} = Z_i(x_1, \ldots, x_n, t) \qquad (i = 1, 2, \ldots, n). \tag{7}$$

Theorem 2. If the null solution of system (7) is uniformly asymptotically stable ([4] p. 32), then, for any $\epsilon > 0$, the positive numbers η and δ can be taken such that at $h < \eta$, the solution $z_i(t)$ of the system (2), defined by the conditions $z_i(0) = z_i^0, |z_i^0| < \delta$, satisfies the inequalities

$$|z_i(t)| < \varepsilon \qquad (i = 1, 2, \ldots, n) \tag{8}$$

over the entire range of $0 \leq t \leq T$ $(0 < T \leq \infty)$.

The theorem 2 is in itself one of the modifications of the theorems of stability at constant disturbances [5]. It is clear that in the present case we speak of the disturbances which, in distinction from [5], are limited by RMS values rather than by mean values. However, the fulfillment of the inequality

$$\int_t^{t+1} \Delta(t)\,dt \leqslant \left(\int_t^{t+1} \Delta^2(t)\,dt \right)^{1/2}, \tag{9}$$

which results from the Buniakowski-Schwarz inequality, shows that the conditions of the theorem 2 provide for the fulfillment of the conditions of the theorem (1.1) in [5]. The validity of the theorem 2 ensues thus from the aforementioned theorem. Let us note, besides, that the theorem 2 can be simply proved in a straightforward manner from the first scheme in work [6]. This proof is given in the appendix in order to provide a rough estimate for the numbers δ, η, which ensure satisfaction of the inequalities (8).

Then, emphasizing that always $H \geq h$, it is possible to make the following formulation.

Theorem 3. If the null solution of the system (7) is uniformly asymptotically stable, then, for any $\epsilon > 0$, the positive numbers η and δ can be taken such that at $H < \eta$, the solution $z_i(t)$ of the system (2), defined by the conditions $z_i(0) = z_i^0$, $|z_i^0| < \delta$, satisfies the inequalities $|z_i(t)| < \epsilon$ $(i = 1, 2, \ldots, n)$ over the entire range of $0 \leq t \leq T$ $(0 < T \leq \infty)$.

Hence, in the case of uniform asymptotic stability of the null solution of system (7), a sufficiently accurate RMS approximation of the $r_i(t)$ functions will provide for satisfactory precision in trajectory approximation.

Finally, we note that if the right sides of the system (7) are independent of t, the asymptotic stability is uniform [7]. If the asymptotic stability of the null solution of system (7) is provided by Lyapunov's function, then this stability is uniform [8].

2. Selection of Optimum Parameter Values

We shall demonstrate on certain particular cases the methods of selecting the control parameters and control functions for which the quantity H has its minimum value.

First, we consider the case where function φ_i has the form

$$\varphi_i = \sum_{k=1}^{m} c_{ik} y_k(t) \quad (i = 1, 2, \ldots, n). \tag{10}$$

The methods of selecting the control functions $y_k(t)$ (task B), as well as both the functions $y_k(t)$ and the parameters c_{ik} (task C) for which H has its minimum value, are given in [1]. In the present work, we solve the problem of determining the optimum system of c_{ik} parameters, i.e., task A.

Obviously, in our case

$$H^2 = \int_0^T \sum_{i=1}^{n} \left(\sum_{k=1}^{m} c_{ik} y_k(t) - r_i(t) \right)^2 dt.$$

A necessary and sufficient condition for obtaining a minimum value of H is to select c_{ik} such that for any i the value of

$$H_i^2 = \int_0^T \left[\sum_{k=1}^{m} c_{ik} y_k(t) - r_i(t) \right]^2 dt$$

be minimum. The validity of this statement follows from the consideration that whatever the selection of the parameters $c_{i1}, \ldots, c_{im}$, it does not affect the quantity H_k, provided $k \neq i$.

In this way, the problem has been reduced to the simplest problem of the theory of root-mean-square approximations. Following the known rule of this theory ([9] p. 396), and assuming the set of functions $y_1(t), \ldots, y_m(t)$ to be linearly independent in the interval $0 \leq t \leq T$, we find the c_{ik} parameters from n sets of the form

$$c_{i1}(y_1, y_j) + \ldots + c_{im}(y_m, y_j) = (r_i, y_j) \quad \begin{pmatrix} i = 1, 2, \ldots, n \\ j = 1, 2, \ldots, m \end{pmatrix}. \tag{11}$$

Here, we introduce the designation $(y_k, y_j) = \int_0^T y_k(t)\, y_j(t)\, dt$.

If we consider the vectors $c_k(c_{1k}, \ldots, c_{nk})$ $(k = 1, 2, \ldots, m)$, then the n sets of the form (11) can be written in the form of a single set as follows

$$c_1(y_1, y_j) + \ldots + c_m(y_m, y_j) = (r, y_j) \quad (j = 1, 2, \ldots, m), \tag{12}$$

where (r, y_j) represents a vector with the projections

$$\int_0^T r_i(t)\, y_j(t)\, dt \qquad (i = 1, 2, \ldots, n).$$

If we find the c_{ik} parameters from the sets (11), then, according to [9], we have

$$H_i^2 = \frac{\Gamma(y_1, \ldots, y_m, r_i)}{\Gamma(y_1, \ldots, y_m)},$$

where both the numerator and denominator contain the Gramm determinants of the corresponding sets of functions.

Hence,

$$H^2 = \sum_{i=1}^{n} H_i^2 = \frac{1}{\Gamma(y_1, \ldots, y_m)} \sum_{i=1}^{n} \Gamma(y_1, \ldots, y_m, r_i). \tag{13}$$

If the set of functions $y_1(t), \ldots, y_m(t)$ is an orthonormalized set in the interval $0 \leq t \leq T$, then from (11) and (12) we get

$$c_{ik} = (r_i, y_k), \quad H_i^2 = (r_i, r_i) - \sum_{k=1}^{m} c_{ik}^2, \quad H^2 = \sum_{i=1}^{n} (r_i, r_i) - \sum_{i=1}^{n} \sum_{k=1}^{m} c_{ik}^2.$$

In the case where the set of functions $y_1(t), \ldots, y_m(t), \ldots$ is a complete set, H can be made arbitrarily small by selecting the number m sufficiently large.

3. Selection of Optimum Initial Conditions

Let us now examine the case where the functions φ_i are given by the formulas

$$\varphi_i = \sum_{j=1}^{m} c_j y_{ij}(t) \qquad (i = 1, 2, \ldots, n). \tag{14}$$

As an example for this case may serve the following problem. Let us examine the system of differential equations

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}, t) + \mathbf{A}(t)\,\mathbf{z}, \quad \frac{d\mathbf{z}}{dt} = \mathbf{B}(t)\,\mathbf{z}, \tag{15}$$

where $\mathbf{x}$ is an n-dimensional and $\mathbf{z}$ an m-dimensional vector, A(t) is a matrix having n rows and m columns, B(t) is a square m-matrix, $f(\mathbf{x}, t)$ is a vector with the projections $f_i(x_1, \ldots, x_n, t)$ $(i = 1, 2, \ldots, n)$.

Let the vector function $\mathbf{x} = \psi(t)$ be given. To be found is a vector function $\mathbf{z}(t)$ defined by the system

$$\frac{d\mathbf{z}}{dt} = \mathbf{B}(t)\,\mathbf{z}, \tag{16}$$

such as to have function $\mathbf{x}(t)$ defined by system (15), approximate the function $\mathbf{x} = \psi(t)$ over a certain segment.

Let $\mathbf{Z}(t)$ be the fundamental matrix of system (16). If $\mathbf{z}(t)$ is the solution of system (16) that is defined by the initial condition $\mathbf{z}(t) = \mathbf{c}$, then we have $\mathbf{z}(t) = \mathbf{Z}(t)\mathbf{Z}^{-1}(t_0)\mathbf{c}$. System (15) now reduces to the system

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}, t) + \mathbf{Y}(t)\,\mathbf{c}, \tag{17}$$

where $\mathbf{Y}(t) = \mathbf{A}(t)\,\mathbf{Z}(t)\,\mathbf{Z}^{-1}(t_0)$.

In this way, the problem evolves down to one of selecting the optimum vector $c(c_1, \ldots, c_m)$. If now we write the system (17) in the scalar form (1), then for the functions φ_i we get the relations (14).

Thus, let φ_i be given by the formula (14); to be found are such values of the parameters $c_1, \ldots, c_m$ at which the quantity

$$H^2 = \int_0^T \sum_{i=1}^n \left[\sum_{j=1}^m c_j y_{ij}(t) - r_i(t)\right]^2 dt = \int_0^T \left[\sum_{j=1}^m c_j y_j(t) - r(t)\right]^2 dt$$

assumes its minimum value. Here, through y_i is denoted the vector with the projections $y_{1j}, y_{2j}, \ldots, y_{nj}$. According to [9] (p. 396), the numbers $c_1, \ldots, c_m$ are to be found from the system

$$c_1(y_1, y_j) + \ldots + c_m(y_m, y_j) = (r, y_j) \quad (j = 1, 2, \ldots, m), \tag{18}$$

where the following designations

$$(y_i, y_j) = \int_0^T y_i(t)\, y_j(t)\, dt, \quad (r, y_j) = \int_0^T r(t)\, y_j(t)\, dt.$$

are introduced.

If $c_1, \ldots, c_m$ have been found according to (18), we have

$$H^2 = \frac{\Gamma(y_1, \ldots, y_m, r)}{\Gamma(y_1, \ldots, y_m)},$$

where both the numerator and denominator contain the Gramm determinants of the corresponding sets of vector functions (or more precisely: the denominator contains the determinant of system (18), while the numerator contains a determinant constructed in an analogous manner.)

If the vectors y_j form an orthonormalized set of functions in the interval $0 \leq t \leq T$, i.e., if $(y_i, y_j) = 0$ at $i \neq j$ and $(y_i, y_i) = 1$, then we have

$$c_j = (r, y_j) \quad (j = 1, 2, \ldots, m), \quad H^2 = \int_0^T r^2(t)\, dt - \sum_{j=1}^m c_i^2.$$

APPENDIX

Proof of theorem 1. Let $x_i(t)$ be a solution of system (7), $z_i(t)$ be a solution of system (2), where $|z_i(t_0) - x_i(t_0)| < \delta$ at $i = 1, 2, \ldots, n$.

Then it is obvious that

$$|z_i(t) - x_i(t)| \leq \delta + \int_{t_0}^t (|Z_i(z_1, \ldots, z_n, t) - Z_i(x_1, \ldots, x_n, t)| + |\varphi_i - r_i|)\, dt.$$

Since, according to (4), $|\varphi_i - r_i| \leq \Delta(t)$, then allowing for the conditions (3), we have

$$|z_i(t) - x_i(t)| \leq \delta + \int_{t_0}^t \left[L \sum_{k=1}^n |z_k(t) - x_k(t)| + \Delta(t)\right] dt.$$

Introducing the function $u(t) = \max_i |z_i(t) - x_i(t)|$, we get $u(t) \leq \delta + \int_{t_0}^t [nLu(t) + \Delta(t)]\, dt$,

thence, by virtue of lemma 1 of [10], it follows that

$$u(t) \leq \delta e^{nL(t-t_0)} + e^{nL(t-t_0)} \int_{t_0}^t e^{-nL(t-t_0)} \Delta(t)\, dt.$$

Applying the Buniakovski-Schwarz inequalities to the integral on the right side, we finally obtain

$$u(t) \leqslant \delta e^{nL(t-t_0)} + \left[\frac{e^{2nL(t-t_0)}-1}{2nL}\right]^{1/2}\left[\int_0^t \Delta^2(t)\,dt\right]^{1/2}. \qquad (19)$$

If we take the null solution of the system (7) as the solution $x_i(t)$, and assume $z = T$, $t_0 = 0$, then we obtain the estimate required.

Proof of theorem 2. According to the definition of the uniform asymptotic stability for the number $\epsilon > 0$, the positive number $\delta < \epsilon$, T may be taken such that from $|x_i(t_0)| < \delta$ $(i = 1, 2, \ldots, n)$ there follows that

$$|x_i(t)| < \frac{\varepsilon}{2} \quad \text{for} \quad t > t_0\,, \quad |x_i(t_0+T_0)| < \frac{\delta}{2}. \qquad (20)$$

T_0 may be assumed an integer without impairing entity.

Let $z_i(t)$ be a solution of system (2) defined by the initial conditions $i_i(t_0) = x_i(t_0)$ $(i = 1, 2, \ldots, n)$.

At $t_0 \leq t \leq t_0 + T_0$, according to (19), there takes place the inequality

$$|z_i(t) - x_i(t)| \leqslant A\left(\int_{t_0}^{t_0+T_0} \Delta^2(t)\,dt\right)^{1/2},$$

where

$$A^2 = \frac{e^{2nLT_0}-1}{2nL}.$$

In this way, accounting for the conditions of the theorem, we have

$$|z_i(t) - x_i(t)| \leqslant AT_0^{1/2}h \quad \text{for} \quad t_0 \leqslant t \leqslant t_0 + T_0.$$

Let $\eta = \frac{1}{2}\delta A^{-1}T_0^{-\frac{1}{2}}$; if $h < \eta$, we have

$$|z_i(t) - x_i(t)| < \frac{\delta}{2} \quad \text{for} \quad t_0 \leqslant t \leqslant t_0 + T_0.$$

From the inequalities (20) and (21) it follows that the solution $z_i(t)$ of system (2), which satisfies the conditions $|z_i(t_0)| < \delta$, satisfies also the conditions

$$|z_i(t)| < \varepsilon \quad \text{for} \quad t_0 \leqslant t \leqslant t_0 + T_0\,, \quad |z_i(t_0+T_0)| < \delta.$$

Taking the number $t_0 + T_0$ as a new initial moment of time, and setting forth analogous considerations, we come to the conclusion that at $t \geq t_0$ we have $|z_i(t)| < \epsilon$ $(i = 1, 2, \ldots, n)$. Obviously, t_0 might have been taken equal to 0 already at the initial point of the discussion.

Annotation. Let the null solution of system (7) be stable according to the exponential law, i.e., let all solutions of system (7) satisfy the inequalities

$$|x_i(t)| < x_0 B e^{-\alpha(t-t_0)} \qquad (i = 1, 2, \ldots, n),$$

where $x_0 = \max_i |x_i(t_0)|$, B, and α are positive constants.

It is readily seen that in proving the theorem 2, it may be assumed that $\delta = \epsilon/2b$ and $T_0 = \ln 2B/\alpha$, on the basis of which A and η are readily computed.

LITERATURE CITED

1. E. A. Barbashin, On the evaluation of the root-mean-square value of the deviation from a given trajectory. Avtomatika i Telemekhanika v. 21, No. 7, 1960.
2. E. A. Barbashin, On the evaluation of the maximum of a deviation from a given trajectory. Avtomatika i Telemekhanika v. 21, No. 10, 1960.

3. E. A. Barbashin, On a certain problem of the theory of dynamic programming. Prikl. matem. i mekh. v. 24, No. 6, 1960.
4. N. N. Krasovskii, Certain problems of the throty of stability of motion. Fizmatgiz, 1959.
5. V. E. Germaidze and N. N. Krasovskii, On stability in the presence of constant disturbances. Prikl. matem. i mekh. v. 21, No. 6, 1957.
6. E. A. Barbashin, On two schemes of proving the theorem of stability in the first approximation. Dokl. AN SSSR. v. 61, No. 1, 1956.
7. I. L. Massera, On Liapounoff's Condition of Stability. Annales of Mathematics vol. 50, No. 3, 1949.
8. I. G. Malkin, Conceming the problem of conversion of Liapounoff's theorem of asymptotic stability. Prikl. matem. i mekh. v. 18, No. 3, 1954.
9. L. Kh. Liberman, On the stability of solutions of integral-differential equations. Izv. vyssh. uchebn. zaved. Matematika No. 3 (4), 1958.

THE ANALYTICAL DESIGN OF CONTROL SYSTEMS

Ya. Kurtsveil'

(Prague)

Translated from Avtomatika i Telemekhanika, Vol. 27, No. 6,
pp. 688-695, June, 1961
Original article submitted January 12, 1961

A mathematical theory is developed for the derivation of an analytical method of designing controls for linear systems with constant coefficients, where this method corresponds to a certain given quadratic-form optimizing functional. The converse problem is also solved.

1. The Direct Problem

We consider a closed control system with generalized coordinates $\eta_1, \eta_2, \ldots, \eta_n$ and control coordinate ξ, described by the equations

$$\dot{\eta}_i = \sum_{j=1}^{n} b_{ij}\eta_j + m_i\xi \quad (i = 1, 2, \ldots, n) \tag{1.1}$$

($b_{i,j}$ and m_i are constants. Let $V(\eta_1, \ldots, \eta_n, \xi) = \sum_{i,j} a_{ij}\eta_i\eta_j + \xi\sum_i a_i\eta_i + \alpha\xi^2$, with $a_{ij} = a_{ji}$ being a given quadratic form. We will solve the following problem (α): for the given initial values $\eta_{10}, \ldots, \eta_{n0}$ we will try to find the functions $\xi(t), \eta_1(t), \ldots, \eta_n(t)$ in the interval $[0, \infty)$, satisfying the equations (1.1), and the conditions

$$\eta_i(0) = \eta_{i0}, \quad \lim_{t\to\infty}\eta_i(t) = 0 \quad (i = 1, 2, \ldots, n) \tag{1.2}$$

and minimizing the integral

$$J(\xi) = \int_0^\infty V[\eta_1(t), \ldots, \eta_n(t), \xi(t)]\,dt. \tag{1.3}$$

(It is assumed that the minimum of the integral $J(\xi)$ is finite.)

Remark. If V is a positive definite form, then the problem (α) is equivalent to the following problem (α'): for given $\eta_{10}, \ldots, \eta_{n0}$ to determine the functions $\xi(t), \eta_1(t), \ldots, \eta_n(t)$ for $t \in [0, \infty)$, such that they satisfy the system (1.1), $\eta_i(0) = \eta_{i0}$ ($i = 1, 2, \ldots, n$) and such that the integral $J(\xi)$ takes its minimum value.

It is obvious that $\lim_{t\to\infty}\inf \sum_{i=1}^{n}\eta_i^2(t) = 0$. Since $J < \infty$, we also have $\int_0^\infty \xi^2(t)\,dt < \infty$.

If $\sum \eta_i^2(t)$ does not tend to zero, then there exists a sequence $t_k \to \infty$ with $\sum \eta_i^2(t_k) = c$, $c < 0$ ($k = 1, 2, 3, \ldots$) and it follows from the formula for the variation of parameters than on some intervals $(t_k - \delta, t_k + \delta)$, $\delta > 0$ ($k = 1, 2, 3, \ldots$), $\sum_i \eta_i^2(t) > c/2$. The integral $J(\xi)$ therefore diverges.

It will sometimes be convenient to use vector notation for the system (1.1) and the form V:

$$\dot{\eta} = B\eta + m\xi,$$
$$V(\eta, \xi) = (A\eta, \eta) + \xi(a, \eta) + \alpha\xi^2. \tag{1.1'}$$

Here $(A\eta, \eta)$ is the scalar product of the vectors $A\eta$ and η.

In the whole of the present article, we will assume that the vectors

$$m, Bm, \ldots, B^{n-1}m \text{ are linearly independent} \tag{1.4}$$

and that

$$V(\eta, \xi) \geqslant \beta\xi^2, \quad \beta > 0 \tag{1.5}$$

for any η and ξ.

We assume that we already know that for any values of $\eta_{10}, \ldots, \eta_{n0}$ the problem (α) has the solution $\xi(t)$, $\eta_1(t), \ldots, \eta_n(t)$. Then, as in [1], from the calculus of variations or from the maximum theorem (see [2]), we can deduce that there exist functions $\lambda_1(t), \ldots, \lambda_n(t)$ such that

$$\xi(t) = \frac{1}{2}\sum_i [m_i\lambda_i(t) - a_i\eta_i(t)], \tag{1.6}$$

$$\dot{\lambda}_i = \sum_j a_{ij}\eta_j + a_i\xi - \sum_j b_{ji}\lambda_i. \tag{1.7}$$

If we substitute (1.6) in (1.1) and (1.7), we obtain

$$\dot{\eta}_i = \sum_j \left(b_{ij} - \frac{m_i a_j}{2\alpha}\right)\eta_j + \sum_j \frac{m_i m_j}{2\alpha}\lambda_j,$$
$$\dot{\lambda}_i = \sum_j \left(a_{ij} - \frac{a_i a_j}{2\alpha}\right)\eta_j + \sum_j \left(-b_{ji} + \frac{m_j a_i}{2\alpha}\right)\lambda_j. \tag{1.8}$$

A. M. Letov in [1] carried the reasoning further, using certain assumptions concerning the matrix of the system (1.8). We will show that these assumptions can be avoided, starting only from the fact that for any values of $\eta_{10}, \ldots, \eta_{n0}$ a solution of the problem (α) exists.

Let X be the set of those points $(\eta_{10}, \ldots, \eta_{n0}, \lambda_{10}, \ldots, \lambda_{n0})$ which are initial points for solutions of the system (1.8) that tend to zero for $t \to \infty$. The set X is evidently a linear subspace in E_{2n}. If the problem (α) has a solution, then there exist values of $\lambda_{10}, \ldots, \lambda_{n0}$, such that the solution $(\ldots, \eta_i(t), \ldots, \lambda_i(t), \ldots)$, $\eta_i(0) = \eta_{i0}$, $\lambda_i(0) = \lambda_{n0}$ of the system (1.8) satisfies the condition $\eta_i(t) \to 0$ for $t \to \infty$ ($i = 1, 2, \ldots, n$). Since $\xi(t)$ satisfies 1.6 and $\int_0^\infty \xi^2(t)dt < \infty$ [see (1.5)], then $\xi(t) \to 0$* and $\sum_j m_j\lambda_j(t) \to 0$ for $t \to \infty$

It is not difficult to establish that if a certain component $\zeta_i(t)$ of the solution of a system of linear differential equations with constant coefficients tends to zero for $t \to \infty$, then $\dot{\zeta}_i(t) \to 0$ for $t \to \infty$.** Since $\sum_j m_j\lambda_j(t)$ is a component of such a system, obtained from (1.8) by using a suitable linear transformation of coordinates, then $\sum_j m_j\lambda_j(t) \to 0$ for $t \to \infty$. We can write the last n equations of (1.8) in the vector form

$$\dot{\lambda} = A\eta - \frac{a}{2\alpha}(a, \eta) - B'\lambda + \frac{a}{2\alpha}(m, \lambda) \tag{1.9}$$

(B' is the matrix transpose of B).

*Since $\ldots, \eta_i(t), \ldots, \lambda_i(t), \ldots$ is a solution of the system (1.8), then $\xi(t) = t^k e^{rt}\sum_{j=1}^{l} \epsilon_j \sin(\sigma_j t + \delta_j) + Z(t)$ where r is real, $k \geq 0$, $\epsilon_j \neq 0$, $\sigma_i \neq \sigma_j$ for $i \neq j$, $\sin\delta_j \neq 0$, if $\sigma_j = 0$ and $t^{-k}e^{-rt}Z(t) \to 0$ for $t \to \infty$. It follows from this that $r < 0$.

**The function $\zeta_i(t)$ can be of similar form to that of $\xi(t)$.

We now form the scalar product of (1.9) with $\underline{m}$

$$(m, \lambda)^{\cdot} = (m, A\eta) - \frac{1}{2\alpha}(m, a)(a, \eta) - (Bm, \lambda) + \frac{1}{2\alpha}(m, a)(m, \lambda)$$

and so $(Bm, \lambda) \to 0$.

We conclude in a similar way that $(Bm, \lambda)^{\cdot} \to 0$ for $t \to \infty$. If is it already known that for some integer $\underline{s}$ we have $(B^s m, \lambda) \to 0$ and $(B^s m, \lambda)^{\cdot} \to 0$ for $t \to \infty$ then when we form the scalar product of (1.9) with $B^s m$ we find that $(B^{s+1} m, \lambda) \to 0$, and, as above, we conclude that $(B^{s+1} m, \lambda)^{\cdot} \to 0$ for $t \to \infty$. It has been proved that $(m, \lambda) \to 0$, $(Bm, \lambda \to 0, \ldots (B^{n-1} m, \lambda) \to 0$ for $t \to \infty$; and since the vectors $m, Bm, \ldots, B^{n-1} m$ form a basis in E_n, then we also have $\lambda(t) \to 0$ for $t \to \infty$.

We have therefore proved the following: if the functions $\eta_1(t), \ldots, \eta_n(t)$ of the set $(\eta_1(t), \ldots, \eta_n(t), \lambda_1(t), \ldots, \lambda_n(t))$, where this set is a solution of the system (1.8), together with the function $\xi(t)$, satisfying (1.6) give a solution of the problem (α), then $(\eta_{10}, \ldots, \eta_{n0}, \ldots, \lambda_{10}, \ldots, \lambda_{n0}) \in X$. Since it has been assumed that for any values $\eta_{10}, \ldots, \eta_{n0}$ the problem (α) has a solution, then it follows from the above reasoning that $\dim X \geq n$ (where $\dim X$ denotes the dimension of X).

We write the matrix Q of the system (1.8) in the form $Q = \begin{pmatrix} Z, & S_1 \\ S_2, & -Z' \end{pmatrix}$, where S_1 and S_2 are symmetric matrices of order $\underline{n}$, Z is a matrix of order $\underline{n}$ and Z' is the transpose of Z. The matrices Q, its transpose Q' and $\begin{pmatrix} Z, & -S_1 \\ -S_2, & -Z' \end{pmatrix}$ have the same invariant polynomials. For the matrices Q' and $\begin{pmatrix} Z, & -S_1 \\ -S_2, & -Z' \end{pmatrix}$ this follows from the following elementary fact: for a given matrix (g_{ij}) $(i, j = 1, 2, \ldots, s)$ and fixed $\underline{k}$ and l we set $h_{ij} = g_{ij}$, if $i \leq k$, $j \leq l$, or $i > k$, $j > l$, and set $h_{ij} = -g_{ij}$ if the converse is true; then

$$\det(h_{ij}) = (-1)^{l-k} \det(g_{ij}).$$

An interchange of rows and columns in the matrix Q' transforms it to the form $\begin{pmatrix} -Z, & S_1 \\ S_2, & Z' \end{pmatrix}$; this means that the matrices $\begin{pmatrix} Z, & -S_1 \\ -S_2, & -Z' \end{pmatrix}$ and $\begin{pmatrix} -Z, & S_1 \\ S_2, & Z' \end{pmatrix}$ have the same invariant polynomials. It follows from this that the elementary divisors of the matrix Q over the complex-number field, relative to the eigenvalues $\rho \neq 0$, occur in pairs $(\lambda - \rho)^k$, $(\lambda + \rho)^k$.

The space E_{2n} is the direct sum of invariant,cyclic spaces I_s (see [3]). Here the minimal polynomial p_s of the space I_s is either a real elementary divisor of the matrix Q or the product of pairs of complex-adjoint elementary divisors. Let ρ_s be a root of the polynomial p_s. The space X is the direct sum of those I_s's for which $\mathrm{Re}\,\rho_s < 0$. Since elementary divisors occur in the pairs $(\lambda - \rho)^k$, $(\lambda + \rho)^k$, then to each space I_s such that $\mathrm{Re}\,\rho_s < 0$ corresponds a space I_{s_1} such that $\mathrm{Re}\,\rho_s < 0$. Consequently, $\dim X = n$ (and there are no eigenvalues for the matrix Q with zero real parts.

Since $\dim X = n$, then for a given $(\eta_{10}, \ldots, \eta_{n0})$, ther is a unique set $(\lambda_{10}, \ldots, \lambda_{n0})$, such that $(\eta_{10}, \ldots, \eta_{n0}, \lambda_{10}, \ldots, \lambda_{n0}) \in X$. Thus the functions $\eta_1(t), \ldots, \eta_n(t)$ from the set $(\eta_1(t), \ldots, \eta_n(t), \lambda_1(t), \ldots, \lambda_n(t))$, when this set yields a solution of (1.8), together with a function $\xi(t)$ satisfying (1.6), form a solution of the problem (α) if and only if $(\eta_{10}, \ldots, \eta_{n0}, \lambda_{10}, \ldots, \lambda_{n0}) \in X$. It thus follows that the solution of the problem (α) is determined uniquely. This can also be proved directly, starting from (1.5).

The mapping $(\eta_{10}, \ldots, \eta_{n0}) \to (\lambda_{10}, \ldots, \lambda_{n0})$, is evidently linear and λ_{i0} is a definite linear function of the variables $\eta_{10}, \ldots, \eta_{n0}$, and so [see (1.6)] $\xi(0) = \sum_j p_j \eta_{j0}$, where the p_j are suitable constants.

Let $\xi(t), \eta_1(t), \ldots, \eta_n(t)$ be a solution of the problem (α) for given $\eta_{10}, \ldots, \eta_{n0}$ and $\tau < 0$. We set $\eta_{i0} = \eta_i(\tau)$. Since the functions $\xi(t+\tau), \eta_1(t+\tau), \ldots, \eta_n(t+\tau)$ give a solution of the problem (α) for given $\eta_{10}, \ldots, \eta_{n0}$, we have

$$\xi(\tau) = \sum_j p_j \eta_j(\tau). \tag{1.10}$$

We thus arrive at the conclusion that the optimal control is given by formula (1.10). If (1.10) is substituted in (1.1), we obtain the system

$$\eta_i = \sum_j (b_{ij} + m_i p_j) \eta_j, \tag{1.11}$$

which has an asymptotically stable, trivial solution.

From the linearity of the mapping of $(\eta_{10}, \ldots, \eta_{n0})$ onto $(\lambda_{10}, \ldots, \lambda_{n0})$, we obtain the following result: let $\xi(t), \eta_1(t), \ldots, \eta_n(t)$ or $\tilde{\xi}(t), \tilde{\eta}_1(t), \ldots, \tilde{\eta}_n(t)$ be solutions of the problem (α) for the initial values $\eta_{10}, \ldots, \eta_{n0}$ or $\tilde{\eta}_{10}, \ldots, \tilde{\eta}_{n0}$; then $\xi(t) + \tilde{\xi}(t)$, $\eta_1(t) + \tilde{\eta}_1(t), \ldots, \eta_n(t) + \tilde{\eta}_n(t)$ also is a solution of the problem (α) for the initial values $\eta_1 + \tilde{\eta}_1, \ldots, \eta_n + \tilde{\eta}_n$.

It remains to determine under what conditions the problem (α) has a solution. We denote by X_0 the algebraic sum of the root subspaces of the matrix B corresponding to the eigenvalues with zero real parts.

Theorem 1. Let the quadratic form $V(\eta_1, \ldots, \eta_n, \xi)$ satisfy (1.5) and let $V(\eta_1, \ldots, \eta_n, \xi) > 0$ if

$$(\eta_1, \ldots, \eta_n) \in X_0, \quad \sum_j \eta_j^2 + \xi^2 > 0. \tag{1.12}$$

Let functions $\xi, \eta_1, \ldots, \eta_n$ exist that satisfy (1.1) and (1.2). Then the problem (α) has a solution.

Proof. After a suitable change of variable, the system (1.1) can be written in the form

$$\dot{y}_1 = B_1 y_1 + l_1 \xi, \quad y_2 = B_2 y_2 + l_2 \xi, \quad y_3 = B_3 y_3 + l_3 \xi. \tag{1.13}$$

Here y_1, l_1; y_2, l_2; y_3, l_3 are vectors; B_1, B_2, B_3 are square matrices having eigenvalues with positive, zero, or negative real parts, respectively. It is obvious that X_0 is the set of those points (y_1, y_2, y_3) for which $y_1 = 0$ and $y_3 = 0$. Let $\xi_n, y_{1n}, y_{2n}, y_{3n}$ $(n = 1, 2, \ldots)$ be an optimizing sequence for the problem (α). $(J(\xi_n)$ tends for $n \to \infty$ to the least possible value for $J(\xi)$. It follows from (1.5) that $\int_0^\infty \xi_n^2(t)\,dt \leqslant k < \infty$. We set $\Xi_n(t) = \int_0^t \xi_n(\tau)\,d\tau$.

From this and from

$$y_{in}(t) = e^{B_i t}\left(y_{in}(0) + \int_0^t e^{-B_i \tau} l_i \xi_n(\tau)\,d\tau\right) = e^{B_i t}\left(y_{in}(0) + \int_0^t e^{-B_i \tau} l_i\,d\Xi_n(\tau)\right) \tag{1.14}$$

it follows that $\Xi_n(t)$, $y_{in}(t)$ are uniformly continuous and bounded functions in every interval $[0, T]$, $T < \infty$. If we extract a suitable sequence, we find that $\Xi_n(t) \to \Xi(t)$, $y_{in}(t) \to y_i(t)$ for $n \to \infty$ uniformly in every interval $[0, T]$.

From a theorem of F. Riesz, it follows that the derivative $\frac{d}{dt}\Xi(t) = \xi(t)$ exists almost everywhere and $\int_0^\infty \xi^2(t)\,dt \leqslant k$ (see [4], Ch. IX). Taking the limit in (1.14)*, we obtain

$$y_i(t) = e^{B_i t}\left(y_i(0) + \int_0^t e^{-B_i \tau} l_i\,d\Xi(\tau)\right) = e^{B_i t}\left(y_i(0) + \int_0^t e^{-B_i \tau} l_i \xi(\tau)\,d\tau\right). \tag{1.15}$$

This means that $\xi(t)$ and $y_i(t)$ satisfy the system (1.13). It is easily verified that

$$\int_0^\infty V\left(y_1(t), y_2(t), y_3(t), \xi(t)\right)dt \leqslant \liminf_{n\to\infty} \int_0^\infty V\left(y_{1n}(t), y_{2n}(t), y_{3n}(t), \xi_n(t)\right)dt.$$

It remains to prove that $y_i(t) \to 0$ for $t \to \infty$.

Since $\|e^{B_i t} x_1\| \geqslant \beta_1 \|x_1\|$ for any x_1 $(\beta_1 > 0, t \geqslant 0)$, then for $y_{1n}(t)$ we have the representation

$$y_{1n}(t) = -\int_t^\infty e^{B_1(t-\tau)} l_1 \xi_n(\tau)\,d\tau. \tag{1.16}$$

Taking the limit in (1.16), we obtain $y_1(t) = -\int_t^\infty e^{B_1(t-\tau)} l_1 \xi(\tau)\,d\tau$ and so $y_1(t) \to 0$, since

* We should note that the variation of the functions $\Xi_n(t)$ in any given interval $[0, T]$, $T < \infty$, is bounded by a constant independent of n.

$\int_0^\infty \xi^2(t)\,dt \leqslant k$. It follows in a similar way from (1.15) that $y_3(t) \to 0$ for $t \to \infty$.

Since $y_1(t) \to 0$ and $y_3(t) \to 0$ for $t \to \infty$, then from (1.12) and from the formula for the variation of parameters we obtain the results that $J(\xi)$ diverges if $y_2(t)$ does not tend to zero. The theorem is thus proved.

We now write the system (1.1) in vector form:

$$\dot{\eta} = B\eta + m\xi.$$

As is known (see [5]), (1.4) guarantees for any initial value η_0 the existence of functions $\xi(t)$, and $\eta(t)$ satisfying (1.14), (1.2), and the condition $J(\xi) < \infty$.

Summing up, we obtain the following theorem.

Theorem 2. Let the conditions (1.4), (1.5), and (1.12) be satisfied. Then the problem (α) has a unique solution $\xi(t)$, $\eta_1(t), \ldots, \eta_n(t)$. This solution satisfies the condition (1.10), where the constants p_j are independent of the initial values $\eta_{10}, \ldots, \eta_{n0}$.

2. The Converse Problem

In this section, we obtain a positive answer to the following question. Let the system (1.11), for suitable values of p_j, possess an asymptotically stable, trivial solution. Then is there a quadratic form V such that (1.10) is a solution of the problem (α)?

We note that the functions $\xi(t) = \sum_j p_j \eta_j(t), \eta_1(t), \ldots, \eta_n(t)$ form a solution of the problem (α) for the equation (1.1), the quadratic form V, and the initial values $\eta_{10}, \ldots, \eta_{n0}$ when and only when the functions $\zeta(t) \equiv 0, \eta_1(t), \ldots, \eta_n(t)$ form a solution of the problem (α) for the system

$$\dot{\eta}_i = \sum_j (b_{ij} + m_i p_j)\eta_j + m_i\zeta \tag{2.1}$$

and the quadratic form $W(\eta_1, \ldots, \eta_n, \zeta) = V\left(\eta_1, \ldots, \eta_n, \sum_j p_j\eta_j + \zeta\right)$ for the same initial values.

We now assume that the trivial solution of the system (1.11) is asymptotically stable. We will use vector notation. We write the system (2.1) in the form

$$\dot{\eta} = D\eta + m\zeta, \quad D = (b_{ij} + m_i p_j), \tag{2.1'}$$

and the form W in the form $W(\eta, \zeta) = (C\eta, \eta) + \zeta(c, \eta) + \gamma\zeta^2$, where $C = (c_{ij})$ $c_{ij} = c_{ji}$.

We will try to find a positive definite form W, such that $\xi \equiv 0$ is a solution of the problem (α) for (2.1) and W, for any initial values. We write the solution of (2.1') in the form

$$\eta(t) = e^{Dt}\eta_0 + \int_0^t e^{D(t-\tau)} m\zeta(\tau)\,d\tau. \tag{2.2}$$

Since $\|e^{Dt}\|^2 = \sum_i \sum_j \left((e^{Dt})_{ij}\right)^2 \leqslant k_1 e^{-\delta t}, \delta > 0,$ the substitution of (2.2) in the form W, after some elementary transformations, yields

$$\int_0^\infty W\left(\eta(t), \zeta(t)\right)dt = \int_0^\infty (Ce^{Dt}\eta_0, e^{Dt}\eta_0)\,dt +$$
$$+ \int_0^\infty \left(c + 2\int_0^\infty e^{D\sigma} Ce^{D\sigma}\,d\sigma m, e^{Dt}\eta_0\right)\zeta(t)\,dt +$$
$$+ \int_0^\infty W\left(\int_0^t e^{D(t-\tau)} m\zeta(\tau)\,d\tau, \zeta(\tau)\right)dt, \tag{2.3}$$

where D' is the transpose of the matrix D. Since W must be positive definite, then $\gamma > 0$, $(C\eta, \eta)$ is a positive definite form, and the relation

$$c = -2\int_0^\infty e^{D'\sigma} C e^{D\sigma}\, d\sigma m \tag{2.4}$$

is a sufficient condition for $\zeta = 0$ to be a solution of the problem for (2.1) and W for any η_0. But (2.4) is also a necessary condition, as can be easily seen by substituting $\epsilon\zeta(t)$ for $\zeta(t)$ and taking the limit for $\epsilon \to 0$. It is evident that W is a positive definite form if $(c, \eta)^2 < 4\gamma (C\eta, \eta)$ for $\eta \neq 0$. We take any positive definite form $(C\eta, \eta)$ and determine c from (2.4). Then there exists $\gamma_0 \geq 0$ such that W is a positive definite form when $\gamma > \gamma_0$. When we carry out the transition from the form W to the form V, we obtain a positive definite form V such that the solutions of the system (1.11) and $\zeta(t) = \sum_j p_j \eta_j(t)$ form a solution of the problem (α). In this way we have obviously obtained all the positive definite forms having the desired properties.

3. The Case of Bounded Rate of Control

Let V be a positive definite form in the variables $\eta_1, \ldots, \eta_n, \xi_1, \xi_2$. We consider the following problem (β): for given initial values $\eta_{10}, \ldots, \eta_{n0}$, and ξ_0, we will try to find functions $\xi(t)$, $\eta_1(t), \ldots, \eta_n(t)$, such that the conditions (1.1), (1.2), $\xi(0) = \xi_0$ are satisfied and the integral $J_2(\xi) = \int_0^\infty V\left(\eta_1(t), \ldots, \eta_n(t), \xi(t), \dot{\xi}(t)\right) dt$.

takes its minimum value. If we set $\eta_{n+1} = \xi$ (see [6]) and $\zeta = \dot{\xi}$, we reduce the problem (β) to the problem (α) for the system

$$\dot{\eta}_i = \sum_{j=1}^n b_{ij}\eta_j + m_i\eta_{n+1}, \qquad \dot{\eta}_{n+1} = \zeta \tag{3.1}$$

and the form $V(\eta_1, \ldots, \eta_n, \eta_{n+1}, \zeta)$ (see the note in section 1).

Let (1.4) be satisfied. We denote by $\bar{B}$ the matrix of the system (3.1) for $\zeta \equiv 0$, $\bar{B} = \begin{pmatrix} B, & m \\ 0, & 0 \end{pmatrix}$ and set $\bar{m} = \begin{pmatrix} 0 \\ \ldots \\ 0 \\ 1 \end{pmatrix}$ Then the vectors

$$\bar{m}, \bar{B}\bar{m} = \begin{pmatrix} m \\ 0 \end{pmatrix}, \quad \bar{B}^2\bar{m} = \begin{pmatrix} Bm \\ 0 \end{pmatrix}, \ldots, \quad \bar{B}^n\bar{m} = \begin{pmatrix} B^{n-1}m \\ 0 \end{pmatrix}$$

are linearly independent. The following theorem follows from Theorem 3.

Theorem 4. If (1.4) is satisfied and if V is a positive definite form, then there is a unique solution for the problem (β). In addition, the equation

$$\dot{\xi} = \sum_{j=1}^n p_j\eta_j + p_{n+1}\xi \tag{3.2}$$

is satisfied (the constants p_j are independent of $\eta_{10}, \ldots, \eta_{n0}, \xi_0$).

As a corollary, we obtain the results that if (1.4) is satisfied, then there are values of $p_j (j = 1, 2, \ldots, n+1)$, such that all the eigenvalues of the matrix

$$\begin{pmatrix} b_{11}, \ldots, b_{1n}, m_1 \\ \ldots\ldots\ldots\ldots \\ b_{n1}, \ldots, b_{nn}, m_n \\ p_1, \ldots, p_n, p_{n+1} \end{pmatrix} \tag{3.3}$$

have negative real parts. It follows from the results of Section 2 of the present article that if all the eigenvalues of the matrix (3.3) have negative real parts, then there exists a positive definite form V such that the solutions of the equations (1.1) and (3.2) are solutions of the problem (β).

LITERATURE CITED

1. A. M. Letov, "The analytical design of control systems," I Avtomatika i Telemekhanika 21, No. 4, 1960.
2. V. G. Boltyanskii, R. V. Gamkrelidze, and L. S. Pontryagin, "The theory of optimal processes," Izv. AN SSSR, 24, No. 1, 1960
3. F. R. Gantmakher, The Theory of Matrices [in Russian] Gostekhizdat, 1954
4. K. P. Natanson, The Theory of Functions of a Real Variable [in Russian] Gostekhizdat, 1957.
5. R. V. Gamkrelidze, "The theory of processes in linear systems optimal in rapidity of action," Izv. AN SSSR 22, 1958.
6. A. F. Filippov, "Some questions in the theory of optimal control," Moscow University News No. 2, 1959.

THE EXTRAPOLATION-GRADIENT METHOD OF SEARCHING FOR THE MINIMUM OF A QUADRATIC FUNCTION

R. É. Vinograd and Yu. V. Geronimus

(Moscow)

Translated from Avtomatika i Telemekhanika, Vol. 22, No. 6,
pp. 696-710, June, 1961
Original article submitted February 18, 1961

In order to minimize a quadratic function in the presence of random noise the authors propose a method involving extrapolation with respect to three values at equally remote points. The steady-state error and the rate of search convergence are computed; the optimal relationships between the search parameters are found.

1. Introduction

In this paper we investigate the operation of an automatic optimizer which searches for the minimum of a quadratic function

$$y(x) = ax^2 + bx + c, \quad a > 0, \tag{1}$$

where a, b, c are the unknown constants when a random error is present at the output of the measuring device. The latter statement means that for a specified argument x the measuring device of the optimizer determines the quantity

$$Y(x) = y(x) + z. \tag{2}$$

rather than the quantity y.

As to the error z, we assume merely that it is identically and independently distributed for each measurement and that its mathematical expectation $\bar{z}$, its dispersion Dz, and its third and fourth order moments are finite.

The search method described below for determining the minimum of y differs from the step and gradient methods [1,2]. Our method uses the idea of extrapolations and is based on the following preliminary concepts.

If the error z were to be absent, then the abscissa of the minimum for (1)

$$x_{\min} = -\frac{b}{2a}$$

could be found by extrapolation with respect to three values y(x) at arbitrary equally remote points $x_0 - h$, x_0, $x_0 + h$ (the number h shall be called the trial step).

In fact, using the substitutions

$$y_0 = y(x_0) = ax_0^2 + bx_0 + c, \qquad y_0^{\pm} = y(x_0 \pm h) = a(x_0 \pm h)^2 + b(x_0 \pm h) + c$$

and specifying the working step by the formula

$$\Delta_0 = \frac{h}{2} \frac{y_0^- - y_0^+}{y_0^- + y_0^+ - 2y_0},$$

it is easy to prove that

$$x_1 \equiv x_0 + \Delta_0 \equiv -\frac{b}{2a}.$$

It is natural to attempt the application of formulas of the same type to the function (2), but in that case the quantity x_1 will be random; thus not only it but its mathematical expectation $\overline{x}_1$ may differ from x_{min}. However, for the natural assumptions made relative to the error z we should expect that $\overline{x}_1$ will lie closer to x_{min} than x_0; if this is so, then it is logical to repeat the process with the same trial step beginning at x_1, etc.

These concepts lead to the following preliminary search scheme.

The point x_0 and the trial step h are chosen arbitrarily, and the transition from x_n to x_{n+1} (which we shall call one cycle) is achieved as follows: we measure the values

$$Y_n^- = Y(x_n - h), \qquad Y_n = Y(x_n) \qquad Y_n^+ = Y(x_n + h),$$

and then compute the working step according to the formula

$$\Delta_n = \frac{h}{2} \frac{Y_n^- - Y_n^+}{Y_n^- + Y_n^+ - 2Y_n} \tag{3}$$

and assume

$$x_{n+1} = x_n + \Delta_n.$$

We find, however, that in this preliminary form the search algorithm is practically useless for the following reason.

The numerator and denominator of the right side of (3) are computed with reference to the function (1) and yield the respective results

$$y_n^- - y_n^+ = -4ahx_n - 2bh, \qquad y_n^- + y_n^+ - 2y_n = 2ah^2, \tag{4}$$

whence it is evident that when there is no error z the denominator is always positive. In the same formula (3) the denominator may prove to be negative or equal to zero due to the random nature of the quantities $Y_n^{\pm}$; this leads to false or impossible values of the working step. Even if we discard the latter values by agreeing to omit one cycle in this case, the search remains supersaturated with random false and excessively large working steps ("hunting" occurs) and computations will show that the process diverges.

In order to avoid this situation, we can choose a certain precautionary number $k > 0$ in advance and agree to perform the working step (3) only in the case where either the denominator of (3) or its absolute value prove to be greater than k. In the contrary case we also have a choice: we either omit the cycle, or perform the working step (3) while replacing the denominator by k. In all there can be four variants of the search scheme.

Investigations similar to those performed below demonstrate that the optimum results are obtained for the following variant:

$$x_{n+1} = x_n + \frac{h}{2} \frac{Y_n^- - Y_n^+}{k}, \qquad \text{if} \quad Y_n^- + Y_n^+ - 2Y_n < k, \tag{5}$$

$$x_{n+1} = x_n + \frac{h}{2}\,\frac{Y_n^- - Y_n^+}{Y_n^- + Y_n^+ - 2Y_n}, \quad \text{if} \quad Y_n^- + Y_n^+ - 2Y_n \geqslant k. \tag{5}$$

It is this variant of the search algorithms which we shall investigate. It is clear that this variant consists of a combination of the extrapolation and gradient methods. The random values of the argument $x_1, x_2, \ldots$ obtained by means of (5) cause the function y to have the random values $y_1, y_2, \ldots$. Moreover, within each cycle we encounter the additional values

$$y_n^{\pm} = y(x_n \pm h).$$

The exact value of the minimum for the function (1) is $y_{min} = c - (b^2/4a)$, and the purpose of the optimization consists of approximating the quantities $y'_n = y_n - y_{min}$, $y'^{\pm}_n = y^{\pm}_n - y_{min}$ to zero in the probability sense as n increases. Since $y'_n, y'^{\pm}_n \geq 0$, an appropriate measure of their over-all deviation from zero, for example, if the arithmetic mean

$$u_n = \frac{1}{3}(y'^-_n + y'_n + y'^+_n).$$

We shall study the following probability problems associated with the series $\{u_n\} = u_1, u_2, \ldots$.

a) computing the mathematic expectation $\bar{u}_n$ and the dispersion Du_n;

b) finding the convergence conditions for the series $\{\bar{u}_n\}$ and $\{Du_n\}$ and computing their limits $U = \lim_{n\to\infty} \bar{u}_n$ and $D = \lim_{n\to\infty} Du_n$;

c) determining the convergence rates for the series;

d) finding the relationships between the characteristics enumerated here for the search quality and their dependence on our arbitrarily chosen parameters – the trial step h and the precautionary number k;

e) determining the allowed requirements relative to the optimalness of the choice of parameters and determining the possibility of realizing these requirements.

The quantity D is called the steady-state dispersion, and the quantities U and $U_0 = \lim_{n\to\infty} \bar{y}'_n$ are called the steady-state search errors. In fact, we are interested only in U, but many details of the investigations are associated with U_0.

The series $\{\bar{u}_n\}$ and $\{Du_n\}$ prove to be convergent at the rate of geometric progressions that have the positive denominators A and L, respectively. The basic part is played by A which shall be called the convergence coefficient. Its closeness to zero denotes rapid convergence, and its closeness to unity denotes slow convergence.

We shall enumerate the final results.

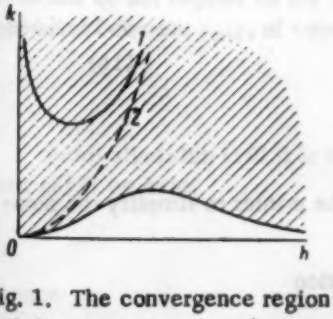


Fig. 1. The convergence region and the optimal curve: 1) optimal curve; 2) $k = 2ah^2$

1. The convergence region for the process is shaded in Fig. 1 and does not contain a series of values for k that are close to zero. Thus the precautionary number cannot be made arbitrarily small (at least not when using the trial step h).

2. A simultaneous reduction of the convergence coefficient A and the steady-state error U is impossible: if h and k are chosen so that $A \to 0$, then $0 \to \infty$; if $U \to 0$, then $A \to 1$.

3. In the convergence region there exists an optimal curve which has the property that from any other point in the region it is possible to move toward the curve in such a way that A decreases and U remains the same, or so that U decreases and A remains the same.

For movement along the optimal curve in the direction of an increase in the trial step h we have $A \to 0$, but $U \to \infty$; in moving in the direction of a decrease in h we have $U \to 0$ but $A \to 1$. The steady-state dispersion D tends towards zero both for $h \to 0$ and $h \to \infty$.

The optimal curve has the shape shown in Fig. 1. Thus, on the optimal curve there exists a positive minimum for k, and k increases without limit both for $h \to 0$ and $h \to \infty$. In other words, it is not practical to perform this pro-

cess with a precautionary number which is too small, and when resorting to large or small trial steps (for accelerating or, on the contrary, for refining the search process) we should in both cases use sufficiently large precautionary numbers.

All the auxiliary computations are cited at the end of the paper in Appendix I and Appendix II.

2. Basic Definitions

Instead of the arbitrary parameters $h > 0$ and $k > 0$ we shall henceforth use the following quantities* that are associated with them in a single-valued manner:

$$\alpha = \frac{1}{2ah^2} \quad \text{and} \quad \beta = \frac{k}{2ah^2} - 1 \tag{6}$$

with a range of variation

$$\alpha > 0, \qquad \beta > -1. \tag{7}$$

We shall reduce the algorithm (5) to a convenient form of notation. Assume z_n, $z_n^{\pm}$ are the values of the error for the measurement of y at the points x_n, $x_n \pm h$, respectively, so that

$$Y_n = y_n + z_n, \qquad Y_n^{\pm} = y_n^{\pm} + z_n^{\pm}.$$

We shall assume that

$$x' = x - x_{\min} = x + \frac{b}{2a}, \qquad y' = y - y_{\min} = y - \left(c - \frac{b^2}{4a}\right)$$

and

$$s_n = \alpha (z_n^- + z_n^+ - 2z_n), \qquad t_n = \sqrt{\frac{\alpha}{8a}} (z_n^- - z_n^+). \tag{8}$$

When these definitions are used we obtain

$$y' = ax'^2 \tag{9}$$

instead of (1) and algorithm (5) becomes

$$\begin{aligned} x'_{n+1} &= \frac{\beta}{1+\beta} x'_n + \frac{t_n}{1+\beta} && \text{for} \quad s_n < \beta, \\ x'_{n+1} &= \frac{s_n}{1+s_n} x'_n + \frac{t_n}{1+s_n} && \text{for} \quad s_n \geqslant \beta \end{aligned} \tag{10}$$

i.e., after obvious transformations which are made on the basis of (4).

Since the previous values of x and y are not encountered again, we shall omit the primes to simplify the notation (i.e., we shall assume $x' = x$, $y' = y$.

Assume the function $g(s)$ is defined for all $s \geq \beta$. We shall introduce the definition

$$[g(s)]_\beta \quad \text{or} \quad g_\beta(s) = \begin{cases} g(\beta) & \text{for} \quad s < \beta, \\ g(s) & \text{for} \quad s \geqslant \beta. \end{cases} \tag{11}$$

Relationships (10) can now be written in simpler form:

* We should remember that a is an unknown, but fully defined positive constant.

$$x_{n+1} = \left[\frac{s_n}{1+s_n}\right]_\beta x_n + \frac{t_n}{[1+s_n]_\beta}. \tag{12}$$

Assuming that the density distribution for the error $\underline{z}$ is specified, it is easy to find (cf. the computations in Appendix I) the density P(S, T) for the joint distribution of the random quantities

$$S_n = z_n^- + z_n^+ - 2z_n, \qquad T_n = z_n^- - z_n^+, \tag{13}$$

which, in view of the propositions governing $\underline{z}$, is independent of $\underline{n}$; then we can also find the density p(s, t) of the joint distribution for s_n and t_n. Here P and $\underline{p}$ are related by a simple function (cf. Appendix I). In particular, if we use the definition

$$\int_{-\infty}^{\infty} T^m P(S, T)\, dT = P_m(S), \qquad \int_{-\infty}^{\infty} t^m p(s, t)\, dt = p_m(s), \tag{14}$$

then

$$p_0(s) = \frac{1}{\alpha} P_0\left(\frac{s}{\alpha}\right), \qquad p_2(s) = \frac{1}{8a} P_2\left(\frac{s}{\alpha}\right), \qquad p_4(s) = \frac{\alpha}{64a^2} P_4\left(\frac{s}{\alpha}\right), \tag{15}$$

these functions are nonnegative and integrable over the entire axis $-\infty < s < +\infty$, and

$$p_1(s) = p_3(s) \equiv 0. \tag{16}$$

It is evident that $p_0(s)$ is the density for s_n, whence

$$\int_{-\infty}^{\infty} p_0(s)\, ds = 1. \tag{17}$$

Then, we shall define the density of the random quantity x_n in terms of $f_n(x)$. From this definition and from (9) it follows that

$$\int_{-\infty}^{\infty} f_n(x)\, dx = 1, \qquad a\int_{-\infty}^{\infty} x^2 f_n(x)\, dx = \overline{y_n}, \qquad a^2\int_{-\infty}^{\infty} x^4 f_n(x)\, dx = \overline{y_n^2}. \tag{18}$$

Since the errors z_n, $z_n^{\pm}$ arise for measurements performed after the value of x_n has already been found and by definition do not depend on the abscissas of the measurements points, it follows that the quantities (8) and x_n are independent (whereas x_{n+1}, of course, depends on x_n, s_n and t_n). Therefore the density of the triplet (x_n, s_n, t_n) will be

$$f_n(x)\, p(s, t). \tag{19}$$

From algorithm (10) it is required to derive the mathematical expectations $\overline{u_n}$ and the dispersions Du_n. On the basis of (9) we have

$$y_n = ax_n^2, \qquad y_n^{\pm} = a(x_n \pm h)^2, \tag{20}$$

whence

$$u_n = \frac{1}{3}(y_n^- + y_n + y_n^+) = y_n + \frac{2ah^2}{3} = y_n + \frac{1}{3\alpha}.$$

This simple relationship permits us in many respects to reduce the study of the series $\{u_n\}$ to a study of the series $\{y_n\}$.

In particular, from this we have

$$u_n = \bar{y}_n + \frac{1}{3\alpha}, \qquad Du_n = Dy_n,$$

and therefore

$$U = U_0 + \frac{1}{3\alpha},$$

and

$$D = \lim_{n\to\infty} Du_n = \lim_{n\to\infty} Dy_n.$$

With respect to computations of the dispersions, we note also that

$$Dy_n = \overline{y_n^2} - (\bar{y}_n)^2,$$

whence

$$D = V_0^2 - U_0^2,$$

where

$$V_0^2 = \lim_{n\to\infty} \overline{y_n^2}.$$

Thus, the problem reduces to computing the quantities $\bar{y}_n$ and $\overline{y_n^2}$ from formula (20) and the algorithm (10) or (12); then

$$U_0 = \lim_{n\to\infty} \bar{y}_n, \qquad V_0^2 = \lim_{n\to\infty} \overline{y_n^2}.$$

and finally

$$U = U_0 + \frac{1}{3\alpha}, \qquad D = V_0^2 - U_0^2.$$

All of the enumerated quantities will evidently depend on the arbitrarily chosen trial step h and precautionary number k; i.e., in terms of the definitions (6) they will depend on the arbitrary parameters α and β. A study of this relationship is a significant portion of our investigation.

3. The Recursion Relationships for the Mathematical Expectations and Dispersions of y_n

From (12) and (20) it follows that

$$y_{n+1} = a\left\{\left[\frac{s_n^2}{(1+s_n)^3}\right]_\beta x_n^2 + 2\left[\frac{s_n}{(1+s_n)^2}\right]_\beta t_n x_n + \frac{t_n^2}{[(1+s_n)^2]}\right\} \equiv \Psi(x_n, s_n, t_n). \qquad (21)$$

Therefore, in view of (19)

$$\bar{y}_{n+1} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Psi(x, s, t)\, f_n(x)\, p(s, t)\, dx\, ds\, dt.$$

Performing the integration while taking (15), (16), (18), and (21) into account, we obtain the recursion relationship for the mathematical expectations $\bar{y}_n$

$$\bar{y}_{n+1} = A\bar{y}_n + B, \qquad (22)$$

where

$$A = \int_{-\infty}^{\infty} \left[\frac{s^2}{(1+s)^3}\right]_\beta p_0(s)\,ds, \tag{23}$$

$$B = a \int_{-\infty}^{\infty} \frac{p_2(s)}{[(1+s)^3]_\beta}\,ds. \tag{24}$$

Analogously, we obtain the relationship for $\overline{y_n^2}$

$$\overline{y_{n+1}^2} = L\overline{y_n^2} + M\bar{y}_n + N, \tag{25}$$

where

$$L = \int_{-\infty}^{\infty} \left[\frac{s^4}{(1+s)^4}\right]_\beta p_0(s)\,ds, \tag{26}$$

$$M = 6a \int_{-\infty}^{\infty} \left[\frac{s^2}{(1+s)^4}\right]_\beta p_2(s)\,ds, \tag{27}$$

$$N = a^2 \int_{-\infty}^{\infty} \frac{p_4(s)\,ds}{[(1+s)^4]_\beta}. \tag{28}$$

The recursion relationships or finite-difference equations (22) and (25) make it possible to compute the mathematical expectations y_n and $\overline{y_n^2}$, and then their limits for $n \to \infty$. In fact, solving Eq. (22) in the conventional manner, we find:

$$\bar{y}_n = \sigma_0 A^n + \frac{B}{1-A} \quad \left(\sigma_0 = \bar{y}_0 - \frac{B}{1-A}\right), \quad \text{if} \quad A \neq 1,$$
$$\bar{y}_n = \bar{y}_0 + nB, \quad \text{if} \quad A = 1. \tag{29}$$

From this it follows from the positive nature of A and B [cf. (23), (24)] that the steady-state error

$$U_0 = \lim_{n\to\infty} \bar{y}_n$$

exists and is finite only when the condition

$$A < 1, \tag{30}$$

is satisfied; in this case

$$U_0 = \frac{B}{1-A} > 0. \tag{31}$$

(The special case $\sigma_0 = 0$ can easily be proven to lead to the same results).

We shall proceed to relationship (25).

We shall substitute the solution (29) which we have found within it while making use of (31); as a result we obtain

$$\overline{y_{n+1}^2} = L\overline{y_n^2} + M\sigma_0 A^n + MU_0 + N.$$

Now it is also possible to solve this equation:

$$\overline{y_n^2} = \tau_0 L^n + M\sigma_0 \frac{A^n - L^n}{A - L} + \frac{MU_0 + N}{1 - L}, \qquad \text{if} \quad L \neq 1,\ L \neq 1,$$

$$\overline{y_n^2} = \tau_0 L^n + M\sigma_0 n L^n + \frac{MU_0 + N}{1 - L}, \qquad \text{if} \quad L = A < 1,$$

$$\overline{y_n^2} = \overline{y_0^2} + M\sigma_0 \frac{1 - A^n}{1 - A} + n(MU_0 + N), \qquad \text{if} \quad L = 1$$

(the case $L = A \geq 1$ is not of interest to us). The initial value τ_0 can easily be determined. We can again prove that the condition bounding* the series $\{\overline{y_n^2}\}$ is the condition

$$L < 1, \tag{32}$$

which also automatically assures its convergence; in this case we have

$$V_0^2 = \lim_{n \to \infty} \overline{y_n^2} = \frac{MU_0 + N}{1 - L}, \tag{33}$$

and therefore the steady-state dispersion is written as

$$D = V_0^2 - U_0^2 = \frac{MU_0 + N}{1 - L} - U_0^2. \tag{34}$$

All of the quantities determined here are functions of the arbitrary parameters α and β; this is directly evident from formulas (23)-(28) if we take (15) into account. Therefore we shall write $A = A(\alpha, \beta)$, $U = U(\alpha, \beta)$ etc.

4. The Allowed and Optimal Values of the Arbitrary Parameters

All of the computations and the missing proofs for this section are cited in Appendix II.

1. The region of allowed parameter values or the region of convergence for the process [satisfaction of inequalities (30) and (32)] is bounded only from below by the curve $\beta = \beta(\alpha)$ lying in the band

$$-1 \leqslant \beta < -\frac{1}{2},$$

where $\beta(\alpha)$ is a nondecreasing function of α and does not coincide identically with -1 (Fig. 2). In view of definition (6) this is result 1 (cf. section 1); Fig. 2 becomes Fig. 1.

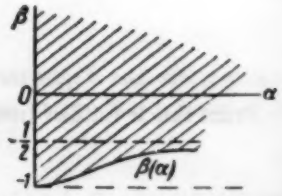


Fig. 2.

2. The level lines for the function $A(\alpha, \beta)$ are arranged in the manner shown in Fig. 3. In particular

a) small values of A are possible only for small α (i.e., from $A \to 0$ it follows that $\alpha \to 0$, and if $\beta > 0$ then it is also true that $\beta \to 0$:

b) moreover, if $\beta \to \infty$, then $A(\alpha, \beta) \to 1$ uniformly with respect to α.

The level lines for $U(\alpha, \beta)$ are arranged as shown in Fig. 4. In particular:

a) if $\alpha \to 0$, then $U(\alpha, \beta) \to \infty$ uniformly with respect to β;

b) small values of $U(\alpha, \beta)$ are possible only for large α; i.e., from $U \to 0$ it follows that $\alpha \to \infty$, and even $\beta \to \infty$ and $\beta/\alpha \to \infty$.

c) each level line U extends at both end in the direction of an unlimited increase of β.

The enumerated facts make it possible to draw a number of conclusions. First of all, result 2 in section 1 derives from them. In fact, from $A \to 0$ it follows that $\alpha \to 0$ and therefore $U \to \infty$; however, if $U \to 0$, then $\beta \to \infty$ and therefore $A \to 1$.

* For the dispersions, it would have been permissible not to require any more than satisfactory search when the other conditions were satisfied.

Thus, a simultaneous decrease of A and U is impossible. Therefore, the problem of involving the optimal choice of the parameters α and β (i.e., the problem involving the optimal values of A and U) can be posed only in the sense of finding the relative minimums of one of these functions for fixed values of the other.

3. When the problem is posed in this manner it cannot be solved, and therefore we come to the concept of an optimal curve in the sense of result 3 in section 1.

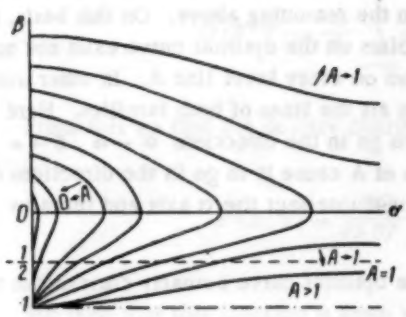


Fig. 3.

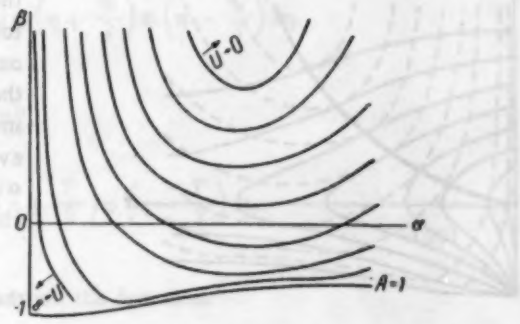


Fig. 4.

In order to prove this we employ the following reasoning.

Irrespective of which of the two functions A and U is assumed fixed, the relative extremum for the other function is found from the equation

$$\begin{vmatrix} \frac{\partial A}{\partial \alpha} & \frac{\partial A}{\partial \beta} \\ \frac{\partial U}{\partial \alpha} & \frac{\partial U}{\partial \beta} \end{vmatrix} = 0, \tag{35}$$

which in expanded form (cf. Appendix II) is written as

$$\alpha \int_{-\infty}^{\beta} p_2(s)\,ds \int_{\beta}^{\infty} \frac{2s^2 p_0(s)\,ds}{(1+s)^3} = \beta \int_{-\infty}^{\beta} p_0(s)\,ds \left[\frac{1-A}{3\alpha} - B + \frac{1}{8} \int_{\beta}^{\infty} \frac{2s}{(1+s)^3} P_2\left(\frac{s}{\alpha}\right) ds \right]. \tag{36}$$

This equation defines a curve on the plane (α, β) which a priori can degenerate into an empty set or can decay into several branches (into the line for relative minimums, relative maximums, etc.). However, none of these branches intersects the α axis, since for $\beta = 0$ the right side of (36) goes to zero and the left side is positive.

In geometric language, Eq. (35) expresses the condition governing the coincidence of the tangents to the level lines for A and U at their common point

$$-\frac{\partial A}{\partial \alpha} \Big/ \frac{\partial A}{\partial \beta} = -\frac{\partial U}{\partial \alpha} \Big/ \frac{\partial U}{\partial \beta}.$$

Therefore, Eq. (35) can be solved graphically by superimposing Fig. 3 onto Fig. 4 and then taking the "line of contacts" for both families. From Fig. 5, it is not difficult to foresee that such a line exists and is precisely the line of relative minimums. Moreover, if the line of contacts were to split up into several branches, then one of them would also be the line of minimums. We shall provide a more thorough justification for this convenient reasoning.

From (35) it is evident that both functions participate equally in the graphical process. We shall take an arbitrary level line $U(\alpha, \beta) = C$. Both of its ends extend in the direction of an unlimited increase of β under conditions

where $A \to 1$, and since $A < 1$ in general in the region under study, it follows that on the line $U = C$ the function A has a minimum point. For continuous variation of C this point describes the curve shown in Fig. 5 (or coincides with one of the branches shown in Fig. 5 in the case where it splits up); this point is optimal in the sense of result 3 in section 1. In fact, by the very graphical process which we have used it is possible to move onto the curve along the level line U from any other point in the region under study and thus decrease A while retaining U constant.

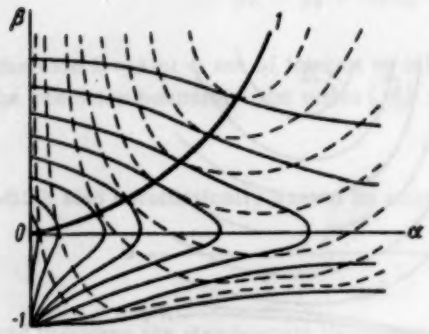


Fig. 5. 1) Optimal curves

In view of the equal weight given to A and U in this process, they can be interchanged in the reasoning above. On this basis, it follows in particular that points on the optimal curve exist not only on each level line U but also on every level line A. In other words, the optimal curve intersects all the lines of both families. Here small values of U cause it to go in the directions $\alpha \to \infty$, $\beta \to \infty$ and even $\beta/\alpha \to \infty$; small values of A cause it to go in the directions of $\alpha \to 0$ and $\beta \to 0$ since it cannot intersect the α axis and remains above that axis.

Thus, the shape of the optimal curve actually corresponds to that shown in Fig. 5.

Assume the equation for the optimal curve is $\beta = \Phi(\alpha)$. It has just been demonstrated that $\Phi(\alpha) > 0$ and $\Phi(\alpha)/\alpha \to \infty$ together with α. Making use of formula (6), we obtain the equation for the optimal curves in the form

$$k = \frac{1 + \Phi(\alpha)}{\alpha} \quad \text{or} \quad k = 2ah^2\left[1 + \Phi\left(\frac{1}{2ah^2}\right)\right]:$$

Now it is clear that $k > 0$, and that for $h \to 0$ or ∞ (i.e., for $\alpha \to \infty$ or 0) we have $k \to \infty$; this is the last postulate in the result 3 of section 1 [relationships (6) cause Fig. 5 to make the transition into Fig. 1].

APPENDIX I

We shall study the random quantities (13). In view of the proposition cited in section 1 relative to the error $\underline{z}$, the quantities z_n, $z_n^{\pm}$, $n = 1, 2, \ldots$ are independent and have the same density distribution $\varphi(\zeta)$; here

$$\int_{-\infty}^{\infty} \zeta^m \varphi(\zeta)\, d\zeta < +\infty \qquad (m = 1.\ 2,\ 3,\ 4). \tag{37}$$

Assume $Q(S, T)$ is the joint distribution for S_n and T_n; i.e., it is the probability for the simultaneous satisfaction of the inequalities

$$z_n^- + z_n^+ - 2z_n = S_n < S, \qquad z_n^- - z_n^+ = T_n < T.$$

In that case

$$Q(S, T) = \iiint_G \varphi(\zeta)\, \varphi(\zeta_1)\, \varphi(\zeta_2)\, d\zeta\, d\zeta_1\, d\zeta_2, \tag{38}$$

where the region G of the space $(\zeta, \zeta_1, \zeta_2)$ is described by the inequalities

$$\zeta_1 + \zeta_2 - 2\zeta < S, \qquad \zeta_1 - \zeta_2 < T.$$

For the substitution of variables

$$\zeta = \eta - \frac{\eta_2}{2}, \qquad \zeta_1 = \eta + \frac{\eta_1}{2}, \qquad \zeta_2 = \eta - \frac{\eta_1}{2}$$

with the Jacobian

$$\frac{\partial(\zeta, \zeta_1, \zeta_2)}{\partial(\eta, \eta_1, \eta_2)} = \frac{1}{2}$$

these inequalities become $\eta_1 < T$, $\eta_2 < S$.

Thus, (38) can be transformed to

$$Q(S, T) = \frac{1}{2} \int_{-\infty}^{T} d\eta_1 \int_{-\infty}^{S} d\eta_2 \int_{-\infty}^{\infty} \varphi\left(\eta - \frac{\eta_2}{2}\right) \varphi\left(\eta + \frac{\eta_1}{2}\right) \varphi\left(\eta - \frac{\eta_1}{2}\right) d\eta.$$

From this we find the density distribution for the pair (S_n, T_n):

$$P(S, T) = \frac{\partial^2 Q}{\partial S\, \partial T} = \frac{1}{2} \int_{-\infty}^{\infty} \varphi\left(\eta - \frac{S}{2}\right) \varphi\left(\eta + \frac{T}{2}\right) \varphi\left(\eta - \frac{T}{2}\right) d\eta. \tag{39}$$

We note that this function is even with respect to T, and therefore the functions

$$P_m(S) = \int_{-\infty}^{\infty} T^m P(S, T)\, dT \tag{40}$$

are identically zero for odd m. However, if m = 2.4, then $P_m(S)$ are integrable over the entire axis since the integrals $\int_{-\infty}^{\infty} P_m(S)\, dS$ can easily be reduced to linear combinations of the integrals (37) using (39) and (40).

Having determined the distribution function Q(S, T) for (13) it is easy to find it for (8) as well. It is obvious that

$$q(s, t) = \mathrm{P}\{s_n < s,\ t_n < t\} = \mathrm{P}\left\{\alpha S_n < s,\ \sqrt{\frac{\alpha}{8a}}\, T_n < t\right\} = Q\left(\frac{s}{\alpha},\ \sqrt{\frac{8a}{\alpha}}\, t\right),$$

where P is the probability symbol.

In this case the density for the pair (s_n, t_n) is

$$p(s, t) = \frac{\partial^2 q}{\partial s\, \partial t} = \frac{\partial^2 Q}{\partial S\, \partial T} \frac{1}{\alpha} \sqrt{\frac{8a}{\alpha}} = \frac{\sqrt{8a}}{\alpha \sqrt{\alpha}} P\left(\frac{s}{\alpha},\ \sqrt{\frac{8a}{\alpha}}\, t\right).$$

This relationship between p(s, t) and P(S, T) leads in obvious fashion to (15) and to (16), since we already know that $P_1(S) = P_2(S) \equiv 0$. Note that if $\varphi(\zeta) \equiv 0$ outside [a, b] then for $P_m(S)$ such a segment will be represented by [-c, c], c = 2(b-a). This can easily be derived from (39) and (40).

APPENDIX II

We must now perform a detailed investigation of the functions (23)-(28). We shall begin with certain general propositions.

Assume that on the semiaxis, $[0, \infty]$, the function g(s) is continuous and has the limit $g(\infty)$, and that $P(S) \geq 0$ is integrable; assume further that $p(s) = (1/\alpha) P(s/\alpha)$. Then

$$\lim_{\alpha \to 0} \int_0^{\infty} g(s)\, p(s)\, ds = g(0) \int_0^{\infty} P(S)\, dS, \quad \lim_{\alpha \to \infty} \int_0^{\infty} g(s)\, p(s)\, ds = g(\infty) \int_0^{\infty} P(S)\, dS.$$

Proof: We have

$$\int_0^\infty g(s)\,p(s)\,ds = \int_0^\infty g(\alpha S)\,P(S)\,dS,$$

where in view of the boundedness of g(s) the latter integral converges uniformly relative to $0 \le \alpha < \infty$; from this the validity of both transitions in the limit follows.

Corollary. If the same conditions are satisfied over the entire axis, then

$$\lim_{\alpha\to 0}\int_{-\infty}^{\infty} g(s)\,p(s)\,ds = (\mu+\nu)\,g(0),$$

$$\lim_{\alpha\to\infty}\int_{-\infty}^{\infty} g(s)\,p(s)\,ds = \mu g(-\infty) + \nu g(\infty),$$

where

$$\mu = \int_{-\infty}^{0} P(S)\,dS, \qquad \nu = \int_0^\infty P(S)\,dS.$$

Assume g(s) is piecewise differentiable and that sg'(s) is bounded; assume further that

$$R(\alpha) = \int_{-\infty}^{\infty} g(s)\,p(s)\,ds, \qquad R_1(\alpha) = \int_{-\infty}^{\infty} g(s)\,P\left(\frac{s}{\alpha}\right)ds.$$

Then

$$R'(\alpha) = \frac{1}{\alpha}\int_{-\infty}^{\infty} sg'(s)\,p(s)\,ds, \tag{41}$$

$$R_1'(\alpha) = \frac{1}{\alpha}\left[R_1(\alpha) + \int_{-\infty}^{\infty} sg'(s)\,P\left(\frac{s}{\alpha}\right)ds\right]. \tag{42}$$

Proof. We have

$$R(\alpha) = \int_{-\infty}^{\infty} g(\alpha S)\,P(S)\,dS.$$

Formal differentiation yields

$$R'(\alpha) = \int_{-\infty}^{\infty} Sq'(\alpha S)\,P(S)\,dS. \tag{43}$$

Since the function $Sg'(\alpha S) = \frac{1}{\alpha}sg'(s)$ is uniformly bounded for the variation of α over any finite segment $0 < \alpha_1 \le \alpha \le \alpha_2$, it follows that the latter integral converges uniformly over this segment; therefore, formal differentiation is valid for any $\alpha > 0$. In order to derive (42) it is sufficient to note that $R_1 = \alpha R$.

Note that in definition (11) we have

$$g_\beta(-\infty) = g(\beta), \qquad g_\beta(\infty) = g(\infty)$$

and

$$\frac{dg_\beta(s)}{ds} = \begin{cases} 0 & \text{for } s < \beta, \\ g'(s) & \text{for } s > \beta. \end{cases}$$

We shall now compute the partial derivatives A, B, L with respect to β and with respect to α. From (23), (24), and (26) we find

$$A'_\beta = \frac{\partial A}{\partial \beta} = \frac{2\beta}{(1+\beta)^3} \int_{-\infty}^{\beta} p_0(s)\, ds,$$

$$B'_\beta = -\frac{2\alpha}{(1+\beta)^3} \int_{-\infty}^{\beta} p_2(s)\, ds,$$

$$L'_\beta = \frac{4\beta^3}{(1+\beta)^5} \int_{-\infty}^{\beta} p_0(s)\, ds.$$

directly.

In order to compute the derivatives with respect to α, we make use of (41) and (42) while taking the definitions (15) into account. As the results we obtain

$$A'_\alpha = \frac{1}{\alpha} \int_{\beta}^{\infty} \frac{2s^3}{(1+s)^5} p_0(s)\, ds,$$

$$B'_\alpha = \frac{1}{\alpha}\left[B - \frac{1}{8} \int_{\beta}^{\infty} \frac{2s}{(1+s)^3} p_2\left(\frac{s}{\alpha}\right) ds\right],$$

$$L'_\alpha = \frac{1}{\alpha} \int_{\beta}^{\infty} \frac{4s^3}{(1+s)^5} p_0(s)\, ds.$$

From this it follows in particular that

$$A'_\alpha > 0 \quad , \quad L'_\alpha > 0,$$

$$A'_\beta \ , \ L'_\beta \text{ have the sign of } \beta. \tag{44}$$

The computation of the derivatives $U = \frac{B}{1-A} + \frac{1}{3\alpha}$ is now obvious.

We establish the fact that

$$A(\alpha, \beta) \leqslant \sqrt{L(\alpha, \beta)}\ , \tag{45}$$

$$\left[\frac{\beta^2}{(1+\beta)^2}\right]_0 = \lim_{\alpha \to 0} A < A(\alpha, \beta) < \lim_{\alpha \to \infty} A = 1 - \mu\left[1 - \frac{\beta^2}{(1+\beta)^2}\right], \tag{46}$$

$$\left[\frac{\beta^4}{(1+\beta)^4}\right]_0 = \lim_{\alpha \to 0} L < L(\alpha, \beta) < \lim_{\alpha \to \infty} L = 1 - \mu\left[1 - \frac{\beta^4}{(1+\beta)^4}\right], \tag{47}$$

where definition (11) is at the left and

$$\mu_1 \frac{\alpha}{(1+\beta)^2} < B(\alpha, \beta) < \mu_2 \frac{\alpha}{(1+\beta)^2}, \tag{48}$$

$$\mu_3 \frac{\alpha}{2\beta+1} < U_0(\alpha, \beta) < \mu_4 \frac{\alpha}{2\beta+1}. \tag{49}$$

are at the right.

In the latter expressions the left sides should be replaced by zeros when $\beta < 0$.

We also establish the fact that

$$M(\alpha, \beta) = \begin{cases} \mu_5 \alpha & \text{for} \quad 0 \leqslant \beta \leqslant 1, \\ \mu_5 \dfrac{\alpha \beta^2}{(1+\beta)^4} & \text{for} \quad \beta > 0; \end{cases} \tag{50}$$

$$N(\alpha, \beta) < \mu_6 \frac{\alpha^2}{(1+\beta)^4}. \tag{51}$$

In formulas (48)-(51), $\mu_i = \text{const} > 0$.

Proof. The expressions for the functions are taken from (23)-(28) while taking (15) into account, and as a rule we shall use definition (11). Note that (23)-(28) are nonnegative. In order to prove (45) it is sufficient to make use of (17) and the Buniakowski inequality

$$A = \int_{-\infty}^{\infty} \left[\frac{s^2}{(1+s)^2}\right]_\beta \sqrt{p_0(s)}\ \sqrt{p_0(s)}\, ds <$$

$$< \sqrt{\int_{-\infty}^{\infty} \left[\frac{s^4}{(1+s)^4}\right]_\beta p_0(s)\, ds}\ \sqrt{\int_{-\infty}^{\infty} p_0(s)\, ds} = \sqrt{L}.$$

Further, if $g(s) = \dfrac{s^2}{(1+s)^2}$, then it is obvious that

$$g_\beta(0) = \left[\frac{\beta^2}{(1+\beta)^2}\right]_0 = \begin{cases} 0 & \text{for} \quad \beta \leqslant 0, \\ \dfrac{\beta^2}{(1+\beta)^2} & \text{for} \quad \beta > 0, \end{cases}$$

and

$$g_\beta(-\infty) = \frac{\beta^2}{(1+\beta)^2}, \qquad g_\beta(\infty) = 1.$$

From this we obtain:

$$\lim_{\alpha \to 0} A(\alpha, \beta) = \left[\frac{\beta^2}{(1+\beta)^2}\right]_0, \qquad \lim_{\alpha \to \infty} A(\alpha, \beta) = \mu \frac{\beta^2}{(1+\beta)^2} + \nu.$$

Since in this case we have

$$\mu + \nu = \int_{-\infty}^{\infty} P_0(S)\, dS = \int_{-\infty}^{\infty} p_0(s)\, ds = 1,$$

in view of (15), (17) and $A(\alpha, \beta)$ increases with respect to α (44), we arrive at (46). In completely analogous fashion, we prove (47). The estimates (48) are obtained directly from (24). It is obvious that

$$\frac{a}{(1+\beta)^2} \int_{-\infty}^{\beta} p_2\, ds < B < \frac{a}{(1+\beta)^2} \int_{-\infty}^{\beta} p_2 ds + \frac{a}{(1+\beta)^2} \int_{\beta}^{\infty} p_2 ds = \frac{a}{(1+\beta)^2} \int_{-\infty}^{\infty} p_2 ds;$$

since $\int_{-\infty}^{\beta} p_2 ds \geqslant \int_{-\infty}^{0} p_2 ds$ for $\beta \geq 0$, and $p_2(s) = \frac{1}{8a} P_2\left(\frac{s}{\alpha}\right)$ (cf. 15), we obtain the relationship (48) where

$$\mu_1 = \frac{1}{8} \int_{-\infty}^{0} P_2(S)\, dS, \quad \mu_2 = \frac{1}{8} \int_{0}^{\infty} P_2(S)\, dS.$$

However, if $\beta < 0$, then we can definitely substitute zero for the lower boundary of B. Now (49) derives from (46) and (48) sihce $U = B/(1-A)$. Inequalities (50) and (51) are proved by the same method as that used to prove (48), but for the function $s^2/(1+s)^4$ we use the coarse estimate $s^2/(1+s)^4 \leq 1$ for $0 \leq s \leq 1$ and the more exact estimate $s^2/(1+s)^4 \leq \beta^2/(1+\beta)^4$ for $s \geq \beta \geq 1$.

We shall now study the level lines A and L. From (44) it is evident that for any level line $A(\alpha, \beta) = C$ we have

$$-\frac{A'_\alpha}{A'_\beta} = \frac{d\beta}{d\alpha} \begin{cases} <0 & \text{for} \quad \beta > 0, \\ =\infty & \text{for} \quad \beta = 0, \\ >0 & \text{for} \quad \beta < 0. \end{cases}$$

From this it follows that any level line or portion of the level line lying above the α axis is a decreasing function of α, and any level line or portion of a level line lying below the α axis is an increasing function of α; if the line lies on the axis proper or intersects it, it has a vertical tangent.

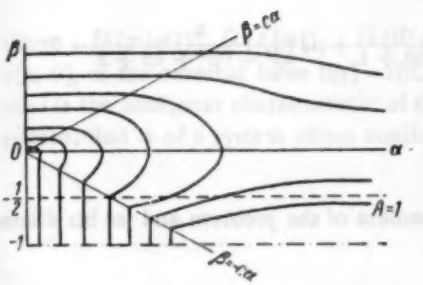


Fig. 6.

The function $A(\alpha, 0)$, in accordance with (44) and (46), increases with respect to α from 0 to $1-\mu$. Therefore for $0 < C < 1-\mu$ the level line $A(\alpha, \beta) = C$ has one and only one common point α_C with the α axis; here $\alpha_C \to 0$ together with C. However, if $1-\mu \leq C < 1$, then no such point exists. The level line adheres to the β axis at the point β_C which is determined from the equation $\left[\frac{\beta^2}{(1+\beta)^2}\right]_0 = C$ [cf (46)]; from this it is evident that β_C also tends toward zero together with C.

The line $A(\alpha, \beta)$ is in a special position.

In accordance with (46) it has only one branch lying below the straight line $\beta = -\frac{1}{2}$. This fully determines the disposition of the curves in Fig. 3.

If $R_m(S) \equiv 0$ outside [-c, c] (cf. Appendix I), then the derivatives (44) may go to zero (for $|\beta| \geq c\alpha$. This has a negligible effect on the remaining reasoning and leads to Fig. 6 instead of Fig. 3.

The investigation of $L(\alpha, \beta)$ is also based on (44) and on the estimate (47) that is fully anologous to (46); therefore it yields the same results. But due to inequality (45) the region $L < 1$ is contained within the region $A < 1$, and this leads to statement 1 in section 4 where the curve $\alpha = \beta(\alpha)$ is the level line $L(\alpha, \beta) = 1$.

We shall study the level lines U. From (49) it follows that for $U = U_0 + \frac{1}{3\alpha}$ we have

$$\mu_3 \frac{\alpha}{2\beta+1} + \frac{1}{3\alpha} < U(\alpha, \beta) < \mu_4 \frac{\alpha}{2\beta+1} + \frac{1}{3\alpha},$$

where we should substitute zero for μ_3 in the left side when $\beta < 0$.

Thus, the level line $U(\alpha, \beta)$ lies between the hyperbolas

$$\mu_3 \frac{\alpha}{2\beta+1} + \frac{1}{3\alpha} = C \qquad \mu_4 \frac{\alpha}{2\beta+1} + \frac{1}{3\alpha} = C, \tag{52}$$

in the region $\beta \geq 0$, and in the region $\beta < 0$ it lies between the straight line $\frac{1}{3\alpha} = C$ and the second of these hyperbolas (and above the line $A = 1$ since $U = \infty$ on the latter).

From Eq. (52) we conclude that as C decreases both hyperbolas (and together with them the entire level line) are shifted in toto to the region of arbitrarily large α and β; this occurs in such a way that $\alpha/\beta \to 0$ (in view of the fact that the terms $\alpha/(2\beta+1)$ go to zero).

We shall estimate the steady-state dispersion. From (34) we have

$$D < \frac{MU_0 + N}{1-L}; \tag{53}$$

On the other hand D is nonnegative as the limit of the dispersions. If we move along the optimal curve to the origin, then L, M, N and U_0 go to zero as is evident from (47)-(51); therefore it is also true that $D \to 0$. If $\alpha \to \infty$ along the optimal curve, then $\alpha/\beta \to 0$ (section 4). In that case we obtain

$$D < \frac{\mu_4\mu_5 \dfrac{\alpha}{2\beta+1}\,\dfrac{\alpha\beta^2}{(1+\beta)^4} + \mu_6 \dfrac{\alpha^2}{(1+\beta)^4}}{\mu\left[1-\dfrac{\beta^4}{(1+\beta)^4}\right]} = \mu_7 \frac{\alpha}{2\beta+1}\,\frac{\alpha\beta^2}{4\beta^3+6\beta^2+4\beta+1} + \mu_8 \frac{\alpha}{4\beta^3+6\beta^2+4\beta+1} \to 0$$

by intensifying inequality (53) by means of (47)-(51).

The authors express their appreciation to A. A. Fel'dbaum for his statement of the problem and for his discussion of the results.

LITERATURE CITED

1. A. A. Fel'dbaum, "The steady-state process in the simplest discrete extremal system where random noise is present," Avtomatika i Telemekhanika 20, 8 (1959).
2. A. A. Fel'dbaum, "The statistical theory of gradient systems for automatic optimization when the object has a quadratic characteristic," Avtomatika i Telemekhanika 21, 2, 1960.

INVESTIGATION OF STABILITY OF PERIODIC STATES IN NONLINEAR PULSE AUTOMATIC SYSTEMS

Ya. Z. Tsypkin

(Moscow)

Translated from Avtomatika i Telemekhanika, Vol. 22, No. 6,
pp. 711-721, June, 1961
Original article submitted December 30, 1960

It is shown that the study of stability of periodic states in nonlinear pulse automatic systems can be reduced to the investigation of a linear pulse system with periodically varying gain or to an equivalent multiconnected linear pulse system with constant parameters. Thus the well-known stability criteria for pulse systems with constant parameters becomes applicable. The effect of random disturbances on periodic states is considered.

The papers [1,2] dealt with the determination of periodic states, a period being a multiple of the sampling period in nonlinear pulse automatic systems (NPAS). The method of harmonic balance forms the basis of these papers. As shown in [2], by operating with discrete periodic functions instead of continuous [1], one is able to obtain the exact solution for periodic states with their half-period $N = 1, 2$, and also an improved approximate solution for periodic states with half-period $N \geq 3$. However, the problem of stability of the obtained periodic states has not been tackled until now. In this paper, an exact solution to this problem is given.

It has been shown that the study of stability of a periodic state can be reduced to the study of stability of the equilibrium position of a linear pulse system with periodically varying gain. The latter system can be made equivalent to a multiconnected linear pulse system with constant coefficients. The transfer functions of the multiconnected pulse system are obtained, to which the well-known stability criteria of ordinary pulse systems are applicable. The effect of random disturbances on periodic states is also considered.

1. Equations of Nonlinear Pulse Automatic System

We shall consider an NPAS consisting of a nonlinear pulse element (NPE) in series with a lumped-parameter continuous part (LCP); their block diagram is shown in Fig. 1. The equation of this system connecting the images

can be written as

$$D\{x[n]\}=D\{f[n]\}-W^*(q)\,D\{\Phi(x[n])\}, \tag{1}$$

where $D\{x[n]\}$, $D\{f[n]\}$, $D\{\Phi(x[n])\}$ denote two-sided discrete Laplace transforms of the error x[n] = x[n, 0], of the external force f[n] = f[n, 0] and of the output quantity from the nonlinear pulse element, respectively; $\Phi(x)$ is the nonlinear characteristic of the pulse element; $W^*(q)$ is the transfer function of the open-loop linear pulse system, that is of a system whose nonlinear pulse element has been replaced by a linear one. The expression

$$D\{x[n]\}=\sum_{n=-\infty}^{\infty} e^{-qn}x[n] \tag{2}$$

denotes the two-sided discrete Laplace transform.*

We assume that a symmetric periodic excitation $\tilde{f}[n]$ of the half-period N is applied to the system. Such a discrete periodic excitation can be represented in the form of a trigonometric polynomial [2]

$$\tilde{f}[n]=\frac{1}{2}\sum_{k=-N_1}^{N_1}{}' c_k e^{j\frac{k\pi}{N}n}. \tag{3}$$

We denote the periodic state response by $\tilde{x}[n]$, where

$$\tilde{x}[n]=\frac{1}{2}\sum_{k=-N_1}^{N_1}{}' a_k e^{j\frac{k\pi}{N}n}. \tag{4}$$

We then obtain for an odd nonlinear characteristic to be used henceforth for the sake of simplicity, the formula

$$\Phi(\tilde{x}[n])=\frac{1}{2}\sum_{k=-N_1}^{N_1}{}' b_k e^{j\frac{k\pi}{N}n}. \tag{5}$$

In the expressions (3)-(5), if N is an odd number $N_1 = N$, but with N even we have $N_1 = N-1$. The dash in the summation sign signified that the summation takes place only over the odd indices [2]. The periodic state equation assumes the form

$$D\{\tilde{x}[n]\}=D\{\tilde{f}[n]\}-W^*(q)\,D\{\Phi(\tilde{x}[n])\}. \tag{6}$$

The determination of symmetric periodic states was presented in [1,2]. Below it shall be assumed that the period state is known.

2. Equations in Terms of Deviations

The ordinary approach is used in the study of the stability of periodic states. Small deviations from the periodic state are considered. If, in due course, the deviation approaches zero, the periodic state is stable; if the deviation increases without bounds, the state is unstable. Equations in terms of deviations shall be formed. It is assumed that at a certain time moment a small disturbance occurs in NPAS due to variation, say, in initial conditions.

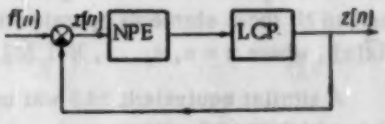


Fig. 1.

* In the case of x[n] ≡ 0 for n< 0, the two-sided discrete Laplace transform reduces to the one-sided one.

This disturbance is fed to the input of a pulse element. Denoting the disturbance by $f_d[n]$, we shall write the external excitation at the input of the pulse element as

$$f[n] = \tilde{f}[n] + f_d[n]. \tag{7}$$

This additional disturbance $f_d[n]$ causes a deviation $\xi[n]$ from the periodic state. Thus

$$x[n] = \tilde{x}[n] + \xi[n]. \tag{8}$$

By substituting (7) and (8) into equation (1), and by subtracting the equation (6) of the periodic state, we obtain the equation for the deviation

$$D\{\xi[n]\} = D\{f_d[n]\} - W^*(q)\, D\{\Phi(\tilde{x}[n] + \xi[n]) - \Phi(\tilde{x}[n])\}. \tag{9}$$

But for small values of the deviation $\xi[n]$ we have

$$\Phi(\tilde{x}[n] + \xi[n]) - \Phi(\tilde{x}[n]) \approx \Phi'(\tilde{x}[n])\,\xi[n],$$

where

$$\Phi'(\tilde{x}[n]) = \left.\frac{d\Phi(x)}{dx}\right|_{x=\tilde{x}[n]}$$

Consequently, the deviation equation becomes a linear one with variable coefficients

$$D\{\xi[n]\} = D\{f_d[n]\} - W^*(q)\, D\{\Phi'(\tilde{x}[n])\,\xi[n]\}. \tag{10}$$

According to the Lyapunov method, the stability of a periodic state is determined by the stability of the system equilibrium position described by an equation in terms of deviation regardless of the neglected terms which correspond to a higher order terms of the expression on the right-hand side of (9).*

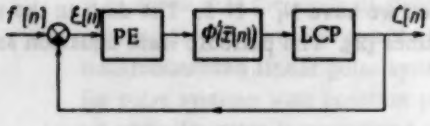


Fig. 2.

But to the equation (10) corresponds a linear PAS with periodically varying gain $k[n] = \Phi'(\tilde{x}[n])$, whose block diagram is shown in Fig. 2. Thus the study of period state in NPAS has been reduced to the study of stability of a PAS with a periodically varying gain (Fig. 2).

3. Equivalent Pulse Automatic Systems

A direct investigation of a PAS with periodically varying gain is not possible because one is unable to find $D\{\xi[n]\}$ from its equation in an explicit form. However, the PAS with variable parameters can be reduced to one with constant parameters. This can be achieved in two ways. The first one is to replace a pulse element with its relative repetition period equal to unity and its variable gain, by a parallel combination of N pulse elements with relative repetition periods equal to N; these elements operate with a shift in relative time equal to unity and possess constant gains equal to $\Phi'(\tilde{x}[n])$, where $n = a, 2, \ldots, N-1$ [3].

A similar equivalent PAS was used by Pyshkin [6] in order to analyze the period state of a PAS with a width-pulse modulation.

* Here the equations are understood to be difference equations; Lyapunov method was generalized to include the latter by E. Cotton [4], P. V. Bromberg [5] and others.

We shall avail ourselves of the second possibility. Consider $\Phi(\tilde{x}[n])$. Since $\Phi(\tilde{x}[n])$ is an odd periodic discrete function, $\Phi'(\tilde{x}[n])$ must be an even periodic discrete function and it can be represented in the form of a trigonometric polynomial

$$\Phi'(\tilde{x}[n]) = \Phi'\left(\frac{1}{2}\sum_{k=-N_1}^{N_1}{}' \dot{a}_k e^{j\frac{k\pi}{N}n}\right) = \frac{1}{2}\sum_{k=-M}^{M}{}' \dot{d}_{2k} e^{j\frac{2k\pi}{N}n}, \tag{11}$$

where M = [N/2] is the integral part of N/2, and $\dot{d}_{2k}$ the complex amplitudes of the harmonics, are equal to

$$\dot{d}_{2k} = d_{2k}e^{j\vartheta_{2k}} = \frac{1}{N}\sum_{\nu=0}^{N-1}{}' \Phi'(\tilde{x}[n])\, e^{-j\frac{2k\pi}{N}\nu} \qquad (|k|<M) \tag{12}$$

and

$$\dot{d}_{\pm 2M} = d_{\pm 2M}e^{j\vartheta_{\pm 2k}} = \frac{1}{N}\sum_{\nu=0}^{N-1}{}' \Phi'(\tilde{x}[n]) e^{\mp j\frac{2M\pi}{N}\nu} \tag{12'}$$

when N is odd, but

$$\dot{d}_{\pm 2M} = d_{\pm 2M} = \frac{1}{2N}\sum_{\nu=0}^{N-1}{}' \Phi'(\tilde{x}[n])\, e^{\mp j n\nu} \tag{12''}$$

when N is even.

In the expressions (11), (12) and (12") the quantities $\dot{d}_0 = d_0$ and $\dot{d}_{\pm 2M} = d_{\pm 2M}$ are real.

By substituting (12) into the equation (10) and by inverting the order of the D-transformation and summation, we obtain

$$D\{\xi[n]\} = D\{f_d[n]\} - \frac{1}{2}W^*(q)\sum_{k=-M_1}^{M_1}{}' \dot{d}_{2k} D\{e^{j\frac{2k\pi}{N}n}\xi[n]\}. \tag{13}$$

By introducing the usual notation for the images

$$\Xi^*(q) = D\{\xi[n]\}, \qquad F_d^*(q) = D\{f[n]\} \tag{14}$$

and noting that, in view of the shift theorem [3]

$$D\{e^{j\frac{2k\pi}{N}n}\xi[n]\} = \Xi^*\left(q - j\frac{2k\pi}{N}\right), \tag{15}$$

we can write the equation (13) as

$$\Xi^*(q) = F_d^*(q) - \frac{1}{2}W^*(q)\sum_{k=-M}^{M}{}' \dot{d}_{2k}\Xi^*\left(q - j\frac{2k\pi}{N}\right). \tag{16}$$

The equation contains the require image $\Xi^*(q)$ for different arguments $(q-j)(2k\pi/N)$. By putting in the equation (16) $(q-j)(2k\pi/N)$ where r = -M, -(M-1), . . . , -1, 0, 1, . . . , M-1, M for an odd N, but r = -M, -(M-1), . . . , -1, 0, 1, . . . , M-1 for an even N, and in view of the periodicity

$$\Xi^*(q \pm j2r\pi) = \Xi^*(q)$$

having the relations

$$\Xi^*\left(q \pm j\frac{2|k|\pi}{N}\right) = \Xi^*\left(q \mp j\frac{2|r|\pi}{N}\right) \quad \text{when} \quad |k|+|r| = N, \tag{17}$$

we obtain a system of equations from which the required images can be determined:

For an odd N, that is N =2M+1

$$\Xi^*\left(q - j\frac{2r\pi}{N}\right) + \frac{1}{2}W^*\left(q - j\frac{2r\pi}{N}\right)\sum_{k=-M}^{M} \dot{d}_{2k}\Xi^*\left(q - j\frac{2(r+k)\pi}{N}\right) = \tag{18}$$

$$= F^*\left(q - j\frac{2r\pi}{N}\right) \qquad (r = -M, -(M-1), \ldots, -1, 0, 1, \ldots, M-1, M);$$

and for an even N = 2M

$$\Xi^*\left(q - j\frac{2r\pi}{N}\right) + \frac{1}{2}W^*\left(q - j\frac{2r\pi}{N}\right)\sum_{k=-(M-1)}^{M-1} \dot{d}_{2k}\Xi^*\left(q - j\frac{2(r+k)\pi}{N}\right) +$$

$$+ \frac{1}{2}W^*(q)\left(\dot{d}_{-2M} + \dot{d}_{2M}\right)\Xi^*\left(q - j\frac{2(M+r)\pi}{N}\right) = F_n^*\left(q - j\frac{2r\pi}{N}\right)$$

$$(r = -M, -(M-1), \ldots, -1, 0, 1, \ldots, M-1).$$

Noting that in agreement with (12*)

$$\frac{\dot{d}_{-2M} + \dot{d}_{2M}}{2} = \frac{d_{-2M} + d_{2M}}{2} = d_{2M}$$

and by denoting

$$d'_{2M} = \frac{d_{2M}}{2},$$

where

$$d'_{2M} = \frac{1}{N}\sum_{\nu=0}^{N-1} \Phi'(\tilde{x}[n])\, e^{-j\pi\nu}$$

is of the same form as the remaining complex amplitudes, we can write the equations in the case of even N as

$$\Xi^*\left(q - j\frac{2r\pi}{N}\right) + \frac{1}{2}W^*\left(q - \frac{2r\pi}{N}\right)\sum_{k=-(M-1)}^{M-1} \dot{d}_{2k}\Xi^*\left(q - j\frac{2(r+k)\pi}{N}\right) +$$

$$+ \frac{1}{2}W(q)\, d'_{2M}\Xi^*\left(q - j\frac{2(M+r)\pi}{N}\right) = F_d^*\left(q - j\frac{2r\pi}{N}\right) \tag{19}$$

$$(r = -M, -(M-1), \ldots, -1, 0, 1, \ldots, M-1).$$

The systems (18) and (19) of equations correspond to equivalent multiconnected linear pulse automatic systems (EMPAS) with a certain number of pulse elements operating synchronously and cophasally.

In view of the shift theorem for the independent variable $\underline{q}$ (see [3]), with the transfer functions $W^*\left(q - j\frac{2r\pi}{N}\right)$ are associated transfer functions $W\left(q - j\frac{2r\pi}{N}\right)$, of continuous units; the block-diagrams of these EMPAS's can be represented in the form as in Fig. 3a, when N is odd, and when N is even as in Fig. 3b.

Solving equations (18) and (19) for $\Xi^*(q)$, we obtain

$$\Xi^*(q) = \sum_{\nu=-M}^{N-M-1} \frac{\Delta^*_{\nu, M+1}(q)}{\Delta^*_N(q)} F_d^*\left(q - j\frac{2r\pi}{N}\right), \tag{20}$$

where for an odd N, N = 2M+1

$$\Delta_N^*(q)=\begin{vmatrix} 1+\frac{d_0}{2}W_M^*, & \dots, & \frac{d_{2M}}{2}W_M^*, & \frac{d_{-2M}}{2}W_M^*, & \dots, & \frac{d_{-2}}{2}W_M^* \\ \frac{d_{-2}}{2}W_{M-1}^*, & \dots, & \frac{d_{2(M-1)}}{2}W_{M-1}^*, & \frac{d_{2M}}{2}W_{M-1}^*, & \dots, & \frac{d_{-4}}{2}W_{M-1}^* \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \frac{d_{-2M}}{2}W_0^*, & \dots, & 1+\frac{d_0}{2}W_0^* & \frac{d_2}{2}W_0^*, & \dots, & \frac{d_{2M}}{2}W_0^* \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \frac{d_4}{2}W_{-(M-1)}^*, & \dots, & \frac{d_{-2(M-1)}}{2}W_{-(M-1)}^*, & \frac{d_{-2M}}{2}W_{-(M-1)}^*, & \dots, & \frac{d_2}{2}W_{-(M-1)}^* \\ \frac{d_2}{2}W_{-M}^*, & \dots, & \frac{d_{-2M}}{2}W_{-M}^*, & \frac{d_{2M}}{2}W_{-M}^*, & \dots, & 1+\frac{d_0}{2}W_{-M}^* \end{vmatrix} \tag{21}$$

and for an even N, N = 2M

$$\Delta_N^*(q)=\begin{vmatrix} 1+\frac{d_0}{2}W_M^*, & \dots, & \frac{d_{2M}}{2}W_M^*, & \dots, & \frac{d_2}{2}W_M^* \\ \frac{d_{-2}}{2}W_{M-1}^*, & \dots, & \frac{d_{2(M-1)}}{2}W_{M-1}^*, & \dots, & \frac{d_{-4}}{2}W_{M-1}^* \\ \dots & \dots & \dots & \dots & \dots \\ \frac{d_{2M}}{2}W_0^*, & \dots, & 1+\frac{d_0}{2}W_0^*, & \dots, & \frac{d_{2(M-1)}}{2}W_0^* \\ \frac{d_2}{2}W_{-(M-1)}^*, & \dots, & \frac{d_{-2(M-1)}}{2}W_{-(M-1)}^*, & \dots, & 1+\frac{d_0}{2}W_{-(M-1)}^* \end{vmatrix} \tag{22}$$

$\Delta_{\nu,M+1}(q)$ are minors obtained by deleting the νth row and $(M+1)$th column from the respective determinant. For the sake of brevity, one abbreviates $W_{\pm k}^*=W^*(q\pm\frac{2k\pi}{N})$. in the formulas (21) and (22).

The concept of a transfer function of an EMPAS is introduced

$$K_{\nu,M+1}^*(q)=\frac{\Delta_{\nu,M+1}^*(q)}{\Delta_N^*(q)}. \tag{23}$$

Then the equation (20) can be finally written as

$$\Xi^*(q)=\sum_{\nu=-M}^{N-M-1}K_{\nu,M+1}^*(q)\,F_d^*\left(q-j\frac{2\nu\pi}{N}\right). \tag{24}$$

We shall now consider particular cases

Where N = 1 we have M = [N/2] = 0, and consequently we obtain from (24)

$$\Xi^*(q)=K_{00}^*(q)\,F_d^*(q), \tag{25}$$

where, in agreement with (23) and (21)

$$K_{00}^*(q)=\frac{1}{1+\frac{d_0}{2}W^*(q)}. \tag{26}$$

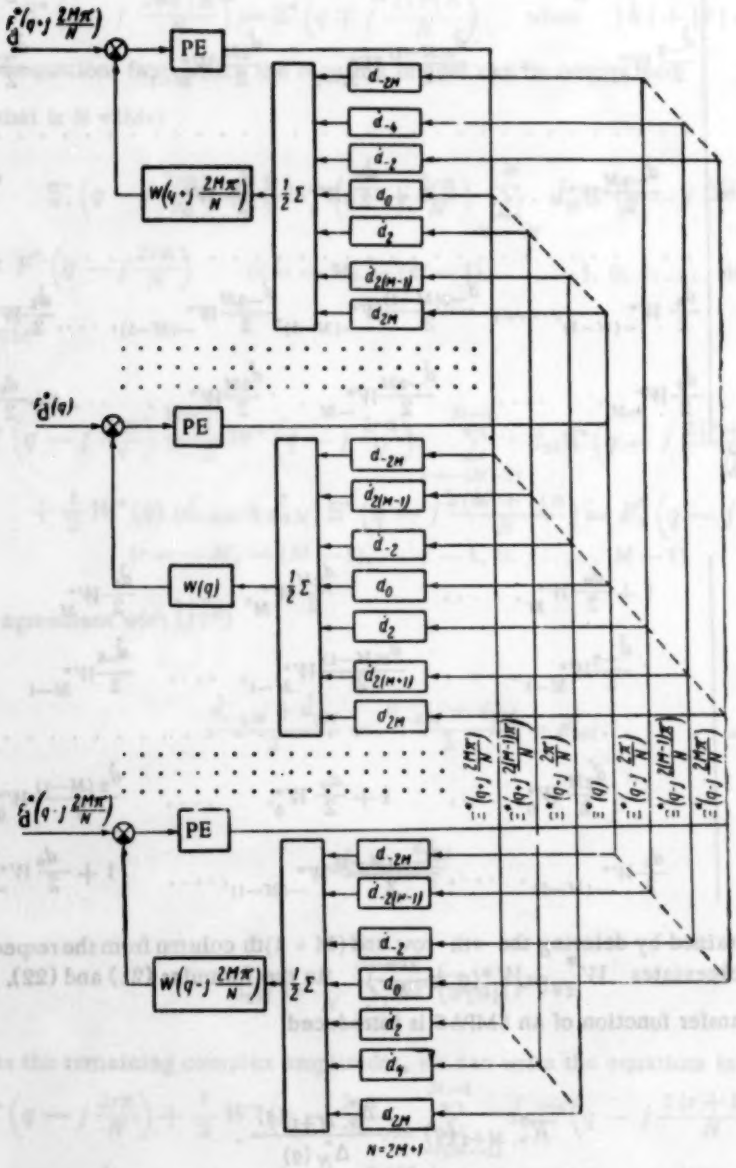


Fig. 3a.

Where N = 2, we have M = [N/2] = 1, and consequently we obtain from (24)

$$\Xi^*(q) = K^*_{-1,2}(q)\, F^*_d\,(q + j\pi) + K^*_{0,2}(q)\, F^*_d\,(q), \tag{27}$$

where in agreement with (24) and (22) we have

$$K^*_{-1,2}(q) = \frac{1 + \frac{d_0}{2} W^*(q)}{\left[1 + \frac{d_0}{2} W^*(q + j\pi)\right]\left[1 + \frac{d_0}{2} W^*(q)\right] + \left(\frac{d_2}{2}\right)^2 W^*(q)\, W^*(q + j\pi)} \tag{28}$$

and $$K^{*}_{-1,0}(q) = \frac{\frac{d_2'}{2} W^{*}(q + j\pi)}{\left[1 + \frac{d_0}{2} W^{*}(q + j\pi)\right]\left[1 + \frac{d_0}{2} W^{*}(q)\right] + \left(\frac{d_2'}{2}\right) W^{*}(q) W^{*}(q + j\pi)} .$$

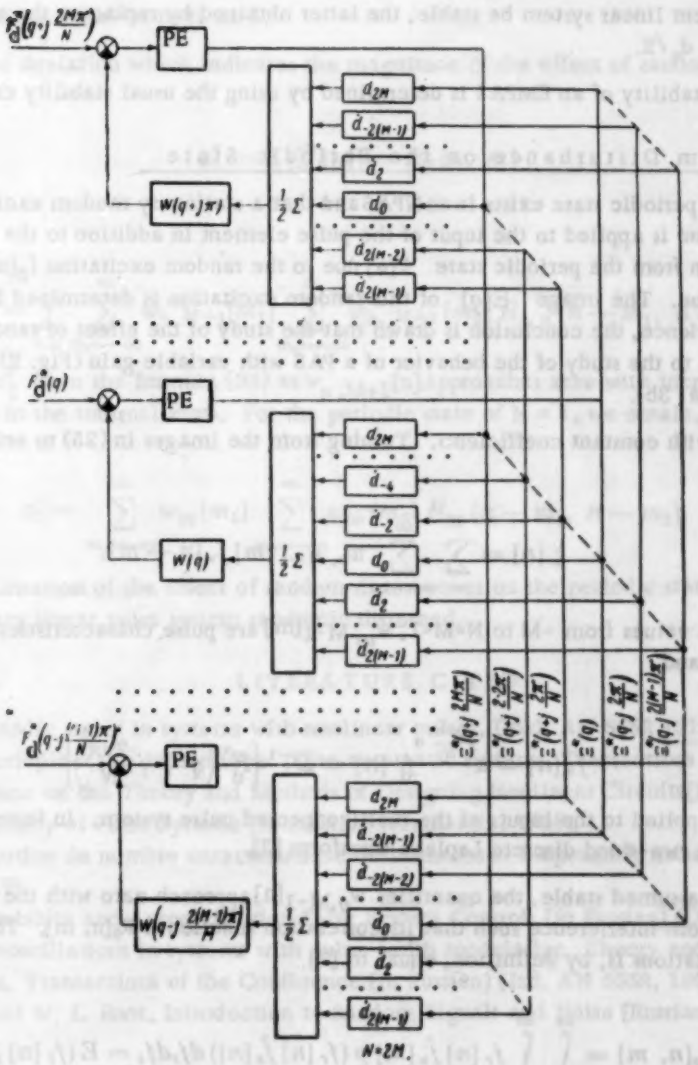


Fig. 3b.

4. Stability Criteria of Periodic States

For the stability of a periodic state it is necessary and sufficient that the deviation $\xi[n]$ from the periodic state approaches zero in the course of time n, that is that the EMPAS be stable.

The stability of an EMPAS, as it is known from the general theory of pulse systems [3], is determined by the position of the poles of its transfer functions $K^{*}_{\nu, M+1}(q)$. If the real parts of the poles are all negative, that is, if the poles are situated in the left half-plane of the complex plane of q, then EMPAS and thus the periodic state of NPAS will be stable.

But the poles of the transfer functions $K^{*}_{\nu M+1}(q)$ coincide with the zeros of their denominators, that is with the roots of the characteristic equation

$$\Delta_N^*(q) = 0. \tag{29}$$

Thus the periodic state in an NPAS is stable, if the real parts of the roots of equation (29) are all negative.

In order to investigate the distribution of the roots of the equation $\Delta_N^*(q) = 0$, the well-known stability criteria of linear pulse systems [3] may be applied.

As seen from (25) and (26), it is necessary and sufficient for the stability of the highest frequency periodic state (N =1), that the equivalent linear system be stable, the latter obtained by replacing the nonlinear pulse element by the linear one with gain $d_0/2$.

When N = 2, the stability of an EMPAS is determined by using the usual stability criteria of linear PAS [3].

5. Effect of Random Disturbance on the Periodic State

We assume that a periodic state exists in an NPAS and that a stationary random excitation $f_d[n]$ sufficiently small in its absolute value is applied to the input of the pulse element in addition to the periodic external excitation. Then the deviation from the periodic state $\xi[n]$ due to the random excitation $f_d[n]$ will also be sufficiently small in its absolute value. The image $\Xi(q)$ of this random excitation is determined from the previously derived equations (10) or (25). Hence, the conclusion is drawn that the study of the effect of random excitations in a periodic state can be reduced to the study of the behavior of a PAS with variable gain (Fig. 2) or of an EMPAS with constant coefficients (Fig. 3a, 3b).

Consider EMPAS with constant coefficients. Turning from the images in (25) to originals, we obtain from the convolution theorem

$$\xi[n] = \sum_{\nu} \sum_{m=-\infty}^{\infty} w_{\nu, M-1}[m] f_\nu[n-m], \tag{30}$$

where ν assumes integral values from -M to N-M-1, $w_{\nu, M-1}[m]$ are pulse characteristics corresponding to the transfer functions $K^*_{\nu, M+1}(q)$ and

$$f_\nu[n] = e^{j\frac{2\nu\pi}{N}n} f_d[n] = D^{-1}\left\{F_d^*\left(q + j\frac{2\nu\pi}{N}\right)\right\} \tag{31}$$

are random excitations applied to the inputs of the multiconnected pulse system. In formula (31) we understand $D^{-1}\{\}$ to be the inverse two-sided discrete Laplace transform [3].

As the system was assumed stable, the quantities $w_{\nu, M-1}[n]$ approach zero with the increase of n. We shall assume that f_{dn} is a random interference such that its correlation function is $R_0[n, m]$. The cross-correlation function of the external excitations is, by definition, equal to [7]

$$R_{r,s}[n, m] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_r[n]\hat{f}_s[m]\, p(f_r[n]\hat{f}_s[n])\, df_r df_s = \mathbf{E}(f_r[n]\hat{f}_s[m]), \tag{32}$$

where $p(f_r\hat{f}_s)$ is the two-dimensional distribution density, and $\hat{f}$ denotes the function conjugate to f.

By substituting the values of $f_\nu[n]$ from (31) into (32) we obtain

$$R_{r,s}[n, m] = e^{j\frac{2\pi}{N}(rn-sm)} R_0[n, m]. \tag{33}$$

We shall find the correlation function of the deviation $\xi[n]$

$$R_\xi[n, m] = \mathbf{E}(\xi[n]\hat{\xi}[m]). \tag{34}$$

By substituting (30) into (34) and by inverting the order of averaging and summation, we obtain

$$R_\xi[n,\ m] = \sum_{\nu,\mu}' \sum_{m_1=-\infty}^{\infty}{}' w_{\nu,\ M+1}[m_1] \sum_{m_2=-\infty}^{\infty} w_{\mu,\ M+1}[m_2]\, \mathrm{E}\,(f[\overline{n}-m_1],\ \hat{f}[\overline{m}-m_2])] \tag{35}$$

or, in view of (32), we finally obtain

$$R_\xi[n, m] = \sum_{\nu,\mu}' \sum_{m_1=-\infty}^{\infty}{}' w_{\nu,\ M+1}[m_1] \sum_{m_2=-\infty}^{\infty} w_{\mu,\ M+1}[m_2]\, R_{\nu,\ \mu}[n-m_1,\ m-m_2], \tag{36}$$

where ν and μ assume values from -M to N-M-1.

The variance of the deviation which indicates the magnitude of the effect of random disturbances on the deviation is

$$\sigma_\xi^2 = \mathrm{E}\,(\xi^2[n]) = R_\xi[n,\ m] \tag{37}$$

or, in view of (36)

$$\sigma_\xi^2 = \sum_{\nu,\mu}' \sum_{m_1=-\infty}^{\infty}{}' w_{\nu,\ M+1}[m_1] \sum_{m_2=-\infty}^{\infty} w_{\mu,\ M+1}[m_2]\, R_{\nu,\ \mu}[n-m_1,\ n-m_2]. \tag{38}$$

When computing σ_ξ^2 from the formula (38) as $w_{\nu,M+1}[n]$ approaches zero with increasing n, one only considers a finite number of terms in the internal sums. For the periodic state of N = 1, we obtain, by putting M = 0 in (38)

$$\sigma_\xi^2 = \sum_{m_1=-\infty}^{\infty}{}' w_{00}[m_1] \sum_{m_2=-\infty}^{\infty} w_{00}[m_2]\, R_{00}[n-m_1,\ n-m_2]. \tag{39}$$

In this case the estimation of the effect of random disturbances on the periodic state is reduced to the study of the behavior of an ordinary linear pulse system randomly disturbed.

LITERATURE CITED

1. M. M. Simkin, Periodic states in systems with nonlinear pulses, Dokl. AN SSSR 131, 6 (1960).
2. Ya. Z. Tsypkin, Periodic states in nonlinear pulse automatic systems, Transactions of the All-Union Interuniversity Conference on the Theory and Methods of Designing Nonlinear Circuits [in Russian] (Tashkent, 1960).
3. Ya. Z. Tsypkin, Theory of Pulse Systems [in Russian] (Fizmatgiz, 1958).
4. E. Cotton, Sur la notion de nombre caracteristique de la methode Liapounoff, Annales de l'Ecole Normale Superieure 36 (1919).
5. P. V. Bromberg, Stability and Autooscillation Pulse System Controls [in Russian] (Oborongiz, 1953).
6. I. V. Pushkin, Autooscillations in systems with pulse-width modulation. Theory and Applications of Discrete Automatic Systems. Transactions of the Conference [in Russian] (Izd. AN SSSR, 1960).
7. W. B. Davenport and W. L. Root, Introduction to Random Signals and Noise [Russian translation] (IL, 1960).

THE STATISTICAL THEORY OF AMPLITUDE QUANTIZATION

A. A. Kosyakin

(Moscow)

Translated from Avtomatika i Telemekhanika, Vol. 22, No. 6,
pp. 722-729, June, 1961
Original article submitted July 13, 1960

This article is devoted to the study of the amplitude quantization of a stationary random signal. It is shown that the quantization process can be considered as a process that adds stationary noise, which we will call quantization noise. We consider conditions under which the quantization noise can be assumed to be practically uncorrelated with the input signal and to have a spectral density that is constant over a very wide frequency range. Methods are developed for applying this theory to automatic control systems.

Automatic control systems containing digital computation devices (DCD) are at the present being applied in a continually widening field. The basic advantages of such systems lie in the possibility of attaining practically unlimited accuracy and interference rejection. The investigation of such systems, however, is very difficult, since in a control system containing DCD the signal, in part of the channel, is in discrete digital form. The representation of a continuous signal in a digital code is essentially nonlinear because of the quantization of the signal amplitude.

There are several articles in the literature devoted to the study of amplitude quantization [1,2]. The main result of these articles is the treatment of the quantization process as one that adds quantization noise, where this noise, under certain conditions, has a practically constant spectral density. These articles, however, do not contain any rigorous theoretical basis for this result; in addition to this, they contain several erroneous conclusions. In particular, the quantization noise is assumed to be independent of the input signal. In the article by Widrow [2], conditions are derived, on the basis of a faulty premise, for which correlation in quantization noise is absent; the resulting conditions appear to be more restrictive than those obtained in the present article. Moreover, the articles referred to consider only the error analysis of quantization in closed coupling channels for pulse-code modulation.

In the present article, we develop a theory for the amplitude quantization of a stationary random signal, and consider methods of applying this theory to the investigation of closed, digital, automatic control systems.

The Problem

The process of amplitude quantization is the nonlinear operation of representing a continuous physical quantity in digital form. The range of the input is divided into a series of adjoining intervals σ_i. All the values of the input that fall in a given interval σ_i are transformed into a certain integer that lies in the interval. A mathematical model of a quantizer is constructed on the basis of the following assumptions.

1. All of the intervals σ_i are of the same length ($\sigma_i = \sigma$) and located so that the ends of the interval σ_i coincide with the inputs $(i-\frac{1}{2})\sigma$ and $(i+\frac{1}{2})\sigma$.

2. If the input to the quantizer lies in the interval σ_i, then the output is $i\sigma$; thus the mean amplification of the quantizer is unity.

3. There is no delay in the quantizer.

Such a quantizer and its characteristics are shown in Fig. 1. Let the quantizer input be the known, two-dimensional probability distribution of a stationary, random process $W_2(x_1, x_2, \tau)$ or the corresponding characteristic function

$$F_{2x}(u_1, u_2, \tau) = M\left[e^{j(u_1x_1+u_2x_2)}\right] = \iint e^{j(u_1x_1+u_2x_2)} W_2(x_1, x_2, \tau)\, dx_1\, dx_2. \tag{1}$$

Then the two-dimensional distribution law for the output of the quantizer will be a distribution law for a two-dimensional random variable taking its values from the set of points with coordinates that are multiples of σ. The probability density of this variable is the sum of terms of the form

$$W_2^*(i\sigma, k\sigma, \tau) = \delta(x_1 - i\sigma, x_2 - k\sigma) \int_{(i-\frac{1}{2})\sigma}^{(i+\frac{1}{2})\sigma} \int_{(k-\frac{1}{2})\sigma}^{(k+\frac{1}{2})\sigma} W_2(x_1, x_2, \tau)\, dx_1\, dx_2, \tag{2}$$

where i and k are integers.

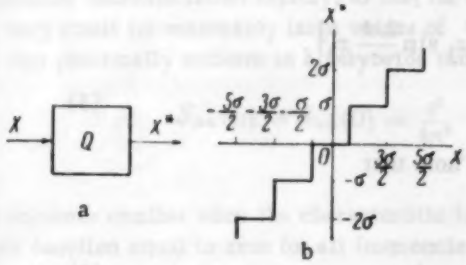


Fig. 1.

As can be seen from (2), the probability density W_2^* is the product of periodically (with period σ) repeating, two-dimensional δ-functions with the function

$$\int_{x_1-\frac{\sigma}{2}}^{x_1+\frac{\sigma}{2}} \int_{x_2-\frac{\sigma}{2}}^{x_2+\frac{\sigma}{2}} W_2(x_1, x_2, \tau)\, dx_1\, dx_2. \tag{3}$$

When we consider the characteristic functions, it is not difficult to obtain

$$F_{2X^*}(u_1, u_2, \tau) = \sum_{i=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} F_{2X}(u_1 + i\varphi, u_2 + k\varphi, \tau) \frac{\sin\frac{\sigma}{2}(u_1 + i\varphi)}{\frac{\sigma}{2}(u_1 + i\varphi)} \frac{\sin\frac{\sigma}{2}(u_2 + k\varphi)}{\frac{\sigma}{2}(u_2 + k\varphi)}, \tag{4}$$

where $\varphi = 2\pi/\sigma$.

Quantization Noise and Its Characteristics

We now consider a quantizer as a combination in parallel of a linear amplifier with amplification coefficient unit and a nonlinear element with the sawtooth -shaped characteristic ψ (Fig. 2)

We will give the name "quantization noise" to the random process defined as the difference between the input and output of the quantizer

$$N = X^* - X. \tag{5}$$

As we can see from Fig. 2, the quantization noise is formed at the output of the nonlinear element with characteristic ψ.

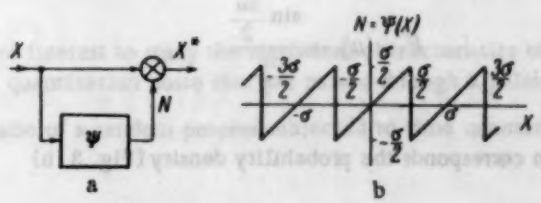


Fig. 2.

In the determination of the statistical characteristics of the quantization noise, we use the expansion of the function $N = \psi(X)$ in a Fourier series

$$N = \Psi(X) = \frac{\sigma}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^n}{n} \sin \frac{2\pi n}{\sigma} X. \tag{6}$$

Then for the autocorrelation functions of the quantization noise $R_{nn}(\tau)$ and for the function $R_{xn}(\tau)$ defining the correlation between the quantization noise and the input signal, we have

$$R_{nn}(\tau) = M[\Psi(x_1)\Psi(x_2)] = \frac{\sigma^2}{\pi^2} \sum_{i=1}^{\infty} \sum_{k=1}^{\infty} \frac{(-1)^{i+k}}{ik} M\left[\sin \frac{2\pi i}{\sigma} x_1 \sin \frac{2\pi k}{\sigma} x_2\right], \tag{7}$$

$$R_{xn}(\tau) = M[x_1 \Psi(x_2)] = \frac{\sigma}{\pi} \sum_{k=1}^{\infty} \frac{(-1)^k}{k} M\left[x_1 \sin \frac{2\pi k}{\sigma} x_2\right]. \tag{8}$$

We now write the sines as differences between exponentials, and note that

$$jM[x_1 e^{ju_2 x_2}] = \frac{\partial F_{2X}(u_1, u_2, \tau)}{\partial u_1}\Big|_{u_1=0}, \tag{9}$$

We finally obtain

$$R_{nn}(\tau) = \frac{\sigma^2}{4\pi^2} \sum_{\substack{i=-\infty \\ i\neq 0}}^{\infty} \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \frac{(-1)^{i+k+1}}{ik} F_{2X}\left(\frac{2\pi i}{\sigma}, \frac{2\pi k}{\sigma}, \tau\right), \tag{10}$$

$$R_{xn}(\tau) = \frac{\sigma}{2\pi} \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \frac{(-1)^{k+1}}{k} \frac{\partial F_{2X}\left(u_1, \frac{2\pi k}{\sigma}, \tau\right)}{\partial u_1}\Big|_{u_1=0}. \tag{11}$$

We assume that the function $F_{2X}(u_1, u_2, \tau)$ is finite, i.e. that it is identically zero everywhere outside the square $|u_1| \leqslant u_{1m}$, $|u_2| \leqslant u_{2m}$. Further, let the conditions*

$$u_{1m} < \frac{2\pi}{\sigma}, \quad u_{2m} < \frac{2\pi}{\sigma}. \tag{12}$$

be satisfied.

Then both the expressions (10) and (11), for the quantization-noise autocorrelation and for the mutual correlation between the quantization noise and the input, become zero. Thus, when the conditions (12) are satisfied, the quantization noise is uncorrelated with the input signal, and is a stationary random process with uncorrelated values. In order to find the mean square quantization noise, we note that if the conditions (12) hold, then the moments of the quantizer output are equal to the sums of the moments of the output and the uncorrelated noise with characteristic function (Fig. 3 a)

$$F_{1N}(u) = \frac{\sin \frac{\sigma u}{2}}{\frac{\sigma u}{2}}. \tag{13}$$

To this characteristic function corresponds the probability density (Fig. 3 b)

* In Widrow's paper [2], the more stringent conditions $u_{1m} \leqslant \frac{\pi}{\sigma}$, $u_{2m} \leqslant \frac{\pi}{\sigma}$. were obtained.

$$W_1(n) = \begin{cases} \frac{1}{\sigma} & \text{for } |n| \leqslant \frac{\sigma}{2}, \\ 0 & \text{for } |n| > \frac{\sigma}{2}. \end{cases} \tag{14}$$

The mean square of the quantization noise is therefore

$$M[n^2] = \int_{-\infty}^{\infty} n^2 W_1(n)\, dn = \frac{\sigma^2}{12}. \tag{15}$$

It should be remarked that the conditions (12) do not hold in any physically realizable case. However, for the majority of probability densities occurring in practice (for example, for the normal distribution), the characteristic functions decrease rather rapidly, so that for sufficiently small steps, the quantization-noise autocorrelation function is very small for reasonably large values of τ. The spectral density of the quantization noise under these conditions is thus practically uniform in a very wide range of frequencies, its value being

$$S_{nn}(\omega) = S_{nn}(0) = \frac{\sigma^2}{4\pi^2} \sum_{\substack{i=-\infty \\ i\neq 0}}^{\infty} \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \frac{(-1)^{i+k+1}}{ik} \int_{-\infty}^{\infty} F_{2X}\left(\frac{2\pi i}{\sigma}, \frac{2\pi k}{\sigma}, \tau\right) d\tau \tag{16}$$

It becomes smaller when the characteristic function of the input signal decreases more rapidly, and in the limit tends to a function equal to zero for all frequencies. However, since the equality

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} S_{nn}(\omega)\, d\omega = R_{nn}(0) = \frac{\sigma^2}{12}, \tag{17}$$

always hold, the spectral density in this limiting case is the singular δ-function on the frequency axis.

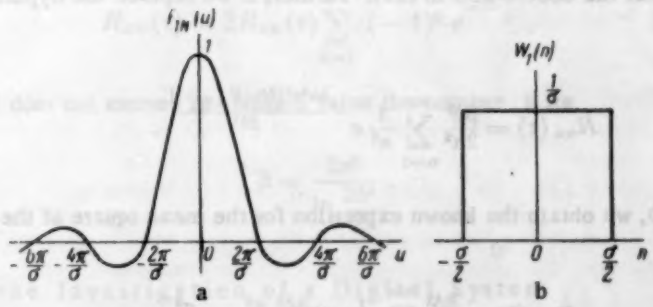


Fig. 3.

Amplitude Quantization in Pulsed Systems

Since the input data and the output results from DCD are not usually produced continuously, but only at definite discrete times, it follows that in systems containing DCD, in addition to the quantization of the signal amplitude, there is also a quantization of the signal in time. The presence of time quantization makes such systems discontinuous, pulse controls.

In this connection, it is of interest to study the statistical characteristics of quantization noise subjected to a double time quantization, i.e. quantization noise that has passed through a pulse element [3].

A convenient characteristic of a random process subjected to time quantization is the discrete spectral density, given by [3]

$$S^*(\omega) = \sum_{m=-\infty}^{\infty} R(mT) e^{-jm\omega T}, \tag{18}$$

where $R(mT) = R(\tau)/_{\tau=mT}$ gives the discrete spectral density of the quantization noise the function

$$S_{nn}^{*}(\omega) = R_{nn}(0) = \frac{\sigma^2}{12}, \tag{19}$$

from which it follows that this discrete spectral density is practically constant over the whole frequency range, and that its value depends only on the size of the steps in the amplitude quantization. This makes it very simple to take amplitude quantization into account in pulse systems.

The Case of a Gaussian Input Distribution

To illustrate what we have said above, we consider the case when the input signal X to the quantizer is a stationary random process with a normal distribution. It is known that the characteristic function of such a process

$$F_{2X}(u_1, u_2, \tau) = \exp\left\{-\frac{1}{2}[(u_1^2 + u_2^2)\delta^2 + 2u_1u_2R_{xx}(\tau)]\right\} \tag{20}$$

is not finite, but it decreases rapidly. If we substitute the characteristic function in the expression (10) for the autocorrelation function of the quantization noise, then after some transformations, we obtain

$$\begin{aligned} R_{nn}(\tau) &= \frac{\beta\delta^2}{\pi^2}\sum_{i=1}^{\infty}\sum_{k=1}^{\infty}\frac{(-1)^{i+k+1}}{ik}e^{-\frac{2\pi^2}{\beta}(i^2+k^2)}\sin h\frac{4\pi^2}{\beta}ikR_{xx}(\tau) = \\ &= \frac{\beta\delta^2}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}e^{-\frac{4n^2\pi^2}{\beta}}\sin h\frac{4n^2\pi^2}{\beta}R_{xx}(\tau) + \\ &+ \frac{\beta\delta^2}{\pi^2}\sum_{i=1}^{\infty}\sum_{k=1}^{\infty}\frac{(-1)^{i+k+1}}{ik}e^{-\frac{2\pi^2}{\beta}(i^2+k^2)}\sin h\frac{4\pi^2}{\beta}ikR_{xx}(\tau), \end{aligned} \tag{21}$$

where $\beta = \sigma^2/\delta^2$ is the depth of the quantization and δ^2 is the mean square of the input signal.

If $\beta \ll 1$, we can neglect the double sum in (21). Further, if we replace the hyperbolic sine by its asymptotic expansion, we obtain*

$$R_{nn}(\tau) = \frac{\beta\delta^2}{2\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}e^{-\frac{4n^2\pi^2[\delta^2 - R_{xx}(\tau)]}{\beta\delta^2}}. \tag{22}$$

In particular, for $\tau = 0$, we obtain the known expression for the mean square of the quantization noise

$$R_{nn}(0) = \frac{\beta\delta^2}{2\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\beta\delta^2}{2\pi^2}\frac{\pi^2}{6} = \frac{\sigma^2}{12}. \tag{23}$$

Let $R_{xx}(\tau) = (\delta^2 \sin\tau\Delta)/\tau\Delta$, corresponding to a uniform distribution in the interval Δ of the spectral density of the input process power. Then the spectral density of the quantization noise will be

$$S_{nn}(\omega) = \frac{4\beta\delta^2}{2\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}\int_0^{\infty}e^{-\frac{4n^2\pi^2}{\beta}\left(1-\frac{\sin\tau\Delta}{\tau\Delta}\right)}\cos\omega\tau\,d\tau. \tag{24}$$

When $(\sin\tau\Delta)/\tau\Delta$ is expanded in a series and only the first two terms are retained (which is permissible since the function $(\sin\tau\Delta)/\tau\Delta$ oscillates and the error is to a large extent cancelled out), we obtain

* This result was first obtained by Bennet in [1]. The further calculations up to (26) are the same as those in the book by B. R. Levin [4].

$$S_{nn}(\omega) = \frac{2\beta\delta^2}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2} \int_0^{\infty} e^{-\frac{2\pi^2 n^2}{3\beta} \tau^2 \Delta^2} \cos \omega\tau \, d\tau \tag{25}$$

or, after integration

$$S_{nn}(\omega) = \frac{\beta\delta^2}{\pi^2\Delta} \sqrt{\frac{3\beta}{2\pi}} \sum_{n=1}^{\infty} \frac{1}{n^3} e^{-\frac{3\beta\omega^2}{8n^2\pi^2\Delta^2}}. \tag{26}$$

It is clear from this formula that the spectral density of the quantization noise can be assumed to be constant up to the frequency

$$\omega_0 = 1{,}9 \frac{\Delta \sqrt{\Delta\gamma}}{\delta\beta \sqrt[4]{\beta}}, \tag{27}$$

where γ is the relative error in percent. The boundary frequency ω_0, as can be seen from the formula (27), increases as the spectrum of the input signal widens and as the depth of quantization β decreases. It should be noted that the condition that β be small is satisfied if the dynamic range of the signal exceeds by several times the size of the quantization step σ.

The spectral density of the quantization noise is given by the expression

$$S_{nn}(\omega) = S_{nn}(0) = \frac{\beta\delta^2}{\pi^2\Delta} \sqrt{\frac{3\beta}{2\pi}} \sum_{n=1}^{\infty} \frac{1}{n^3} = \frac{\sigma^2 \sqrt{\beta}}{12\Delta}. \tag{28}$$

The mutual correlation between the quantization noise and the input signal is obtained from (11) and (20) after some transformation, and is

$$R_{xn}(\tau) = 2R_{xx}(\tau) \sum_{k=1}^{\infty} (-1)^k e^{-\frac{2\pi^2 k^2}{\beta}}, \tag{29}$$

It therefore follows that it does not exceed in absolute value the number ξ for

$$\beta = \frac{2\pi^2}{\ln \frac{2\delta^2}{\xi}}. \tag{30}$$

An Application to the Investigation of a Digital System

We consider the digital-system scheme shown in Fig. 4. Here CD is the converter for transforming continuous quantities into digital form, the block diagram of which can be given in the form shown in Fig. 5.

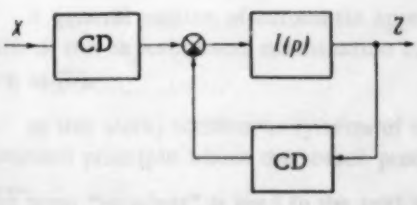
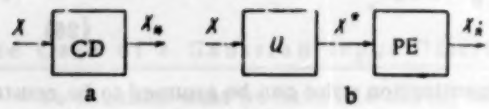


Fig. 4.

Following Ya. Z. Tsypkin in [5], we use the term "limiting pulse system" for a system differing from a digital system in the absence of amplitude quantization. Then if the quantization steps are small enough, the digital system in Fig. 4 becomes a limit pulse system. Acting on this system is not only the useful signal, but also two noise components, uncorrelated with the input signal, and due to the amplitude quantization in the input circuit and in the feedback circuit. The discrete spectral density of each of these noise components is constant and equal to $\frac{1}{12}\sigma^2$. In the investigation of this system, all the methods for the analysis and synthesis of pulse systems under the action of stationary random perturbations were used.

Now, let some interference as well as the useful signal act at the input. It is not difficult to see that the addition of this interference can only broaden the dynamic range of the input signal. This means that the spectral density of the quantization noise remains unchanged, and that the error caused by the action of the interference on the digital system is the same as the error that would be caused the the action of the same interference on a limit pulse system.



a b

Fig. 5.

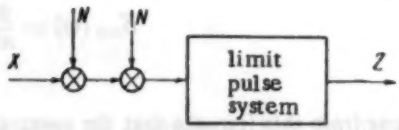


Fig. 6.

SUMMARY

1. The amplitude-quantization process of a stationary, random, input signal can be considered as the process of adding to the input signal a certain equivalent noise, called the quantization noise [1,2].

2. The autocorrelation function of the quantization noise and the mutual correlation between the quantization noise and the input signal can be determined from the formulas (10) and (11) by the known methods of characteristic functions or from the probability density of the input signal.

3. If the amplitude quantization steps are small enough in comparison with the mean square input signal, then the quantization noise can be assumed to be uncorrelated with the input signal. The exact conditions for the validity of this result are given by the inequalities (12).

4. With the same condition that the quantization steps be small, the spectral density of the quantization noise can be assumed to be constant over a wide range of frequencies, and its magnitude can be determined from (16). The discrete spectral density of the quantization noise depends only on the size of the quantization steps, and is equal to $^1/_{12}\sigma^2$.

5. For sufficiently small amplitude-quantization steps, a digital system reduces to a limit pulse system, at the input of which acts not only the useful signal, but also the quantization noise with spectral density $^1/_{12}\sigma^2$.

6. If there is interference acting at the input of a digital system as well as the useful signal, then the error caused by this interference is equal to the error that would occur in a limit pulse system under the influence of the same interference.

LITERATURE CITED

1. W. R. Bennet, Spectra of Quantized Signals, Bell System Technical Journal vol. 27, July, 1948.
2. B. A. Widrow, Study of Rough Amplitude Quantisation by means of Niquist Sampling Theory. Trans. IRE vol. PGCT-3. No. 4, December, 1956.
3. Ya. Z. Tsypkin, The Theory of Pulse Systems [in Russian] Fizmatgiz, 1958.
4. B. R. Levin, The Theoryof Random Processes and Their Application in Radio-Technology [in Russian] Izd. "Sov. radio," 1957
5. Ya. Z. Tsypkin, "An Estimate of the Influence of Amplitude Quantization on Processes in Digital Automatic Control Systems," Avtomatika i telemekhanika 21, Vol. 3, 1960.

EXTREMAL RECEPTION OF SIGNALS

A. A. Krasovskii

(Moscow)

Translated from Avtomatika i Telemekhanika, Vol. 22, No. 6,
pp. 730-738, June, 1961
Original article submitted August 30, 1960

Systems are considered in which the received signal is compared with the signal from an automatically adjustable internal generator. The adjustment of the generator is carried out by using an external control system which operates on the minimizing principle of a criterion of differences between the generated signal and the received one. The equations are obtained and the dynamic process is analyzed in a continuous system of this type.

Receivers of various kinds of signals* usually belong to systems requiring some kind of adjusting operation. Hence it seems natural to apply the principles of automatic self-adjustment to signal receivers.

There are various possible ways of implementing a self-adjusting receiver. As in the case of ordinary hand adjustment, it is possible to have an automatic adjustment of passive networks of the receiver based on the principle of reaching the extremum of certain indicators of the quality of reception.

In this paper an account is given of another principle of attaining extremal reception which is as follows.

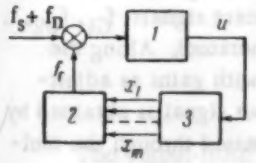


Fig. 1.

The received signal contains a useful component f_s and interference f_n. In the general case, the noise of interference can be assumed to be an arbitrary random time function. Only at the synthesis stage of an optimum self-adjusting system is it necessary that the statistical characteristics of the interference f_n be known. The general form of the useful component f_s is assumed to be known beforehand; it is characterized by a number of parameters x_{si} such as amplitude, carrier frequency, duration and phase of the pulse train, etc. Thus $f_s = f_s(t, x_{s1}, x_{s2}, \ldots x_{sm})$.

The initial information on the useful component of the received signal gives the composition of the parameters x_s only; the accurate values of these parameters at a given moment of time are not known.** The received signal is compared with the signal f_G from an automatically adjustable internal generator (Fig. 1).

The construction of the internal generator is such that the form of the signal $f_G = f_G(t, x_1, x_2, \ldots x_m)$ as well as the adjustment parameters x_i of the generator correspond, at least approximately, to the form and the parameter construction of the signal useful component so, that the differences $f_s - f_G$ can be made sufficiently small by selecting the values of the parameters x_i within the admissible range of f_s.

The difference between the received and the generated signal proceeds to the deviation estimator, block 1. The block output is a criterion of the current deviation of the generated from the received signal. This quantity is fed to the optimizer 3, a system of extremal regulation which implements the search and leads to an adjustment of the generator 2 such that a deviations criterion is minimum. The output signal of the system of extremum reception is either the signal of the adjusted generator or only the adjustment parameters x_i.

A general outline of automatic approximation using a self-adjusting device was briefly developed in [4]. The results of the experimental examination of one case of approximation with the aid of an automatic optimizer were given in [5].

In this work, continuous systems of extremal reception are investigated, the study being based on the above mentioned principle where the search proceeds by the gradient method.

*The term "receiver" is used in the text in its wider meaning and indicates any receptor of information.

**A signal with all the parameters fully given does not carry any information.

Block Diagrams and Equations of the System

The block diagram of a continuous system of extremal reception is shown in Fig. 2. Insofar as the search for the extremum is concerned, the system is in every way analogous to the multidimensional continuous systems of extremal regulation with synchronous detection [1-3].

The search oscillations δx_i correspond to the adjustment parameters of the generator 2. If the extremum of the criterion is not attained, these oscillations produce synchronous variations in the magnitude of the criterion. The latter are extracted by means of synchronous detectors, that is, amplifiers to which the search signals are sent as reference signals. These signals thus obtained pass through filters W_{fi} and integrators, and produce variations in the generator adjustments as long as the synchronous components of oscillation appear of the deviations criterion, that is, until the extremum of this criterion is attained.

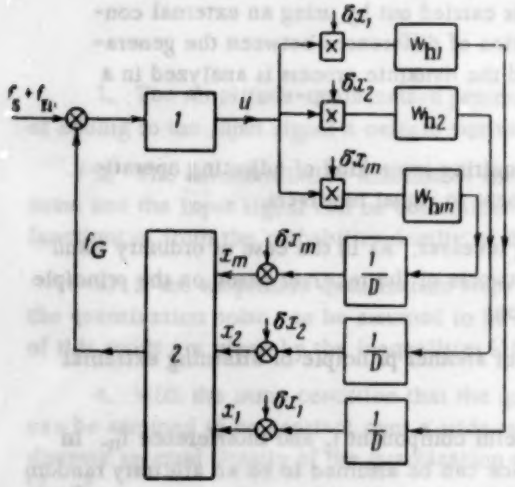


Fig. 2.

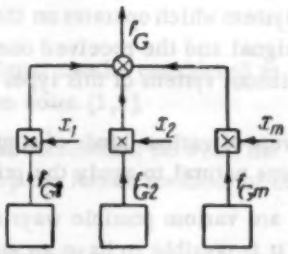


Fig. 3.

One can cite the scheme shown in Fig. 3 as an example of an adjusting generator. Here, there are generators with constant adjustment producing m constant signals $f_{G1}, f_{G2}, \ldots, f_{Gm}$ ("carrier" functions of the generator). Along the signal paths one finds the multipliers with gains as adjustment parameters. The generator output signal is obtained by summing the signals after they have passed through the multipliers. Thus the generator signal becomes in this case a linear function of the adjustment parameters

$$f_G(t, x_1, x_2, \ldots, x_m) = \sum_{i=1}^{m} x_i f_{Gi}(t). \tag{1}$$

Before considering the equations corresponding to the system, one should note that different functions may be taken as the deviations criterion. The use of a squaring device as a deviation block driterion corresponds to a quadratic function

$$u = (f_s + f_n - f_G)^2. \tag{2}$$

The use of a full-wave rectified linear detector as a deviation block criterion corresponds to the absolute value of deviations

$$u = |f_s + f_n - f_G|. \tag{3}$$

In all cases, the deviation criterion can be represented in the form

$$\begin{aligned} u = F[f_s(t, x_{s1}, \ldots, x_{sm}) + f_n(t) - f_G(t, x_1, \ldots, x_m)] = \\ = F[f_n(t)] + 2f_n(t)\sum_{i=1}^{m} f_i(t)\Delta x_i + \sum_{i,j=1}^{m} f_{ij}(t)\Delta x_i \Delta x_j \\ + \Phi(t, x_{s1}, \ldots, x_{sm}, x_1, \ldots, x_m). \end{aligned} \tag{4}$$

Here $\Delta x_i = x_{si} - x_i$ are the deviations of the signal parameters of the generators from those of the useful component of the received signal; the functions $f_i(t)$, $f_{ij}(t) = f_{ji}(t)$ depending on time are close to the functions f_s, f_G.

For many deviation criteria (that is, for many functions F) and for a number of functions f_s and f_G, the term $\Phi(t, x_{c1}, \ldots, x_{cm})$ in the expression (4) is of a higher order of smallness in respect of Δx_i than the preceding terms. Thus, for example, if the adjustment (modulation) of the generator is used, expressed by the formula (1), with an analogous modulation of the useful signal

$$f_s(t, x_{s1}, \ldots, x_{sm}) = \sum_{i=1}^{m} x_{si} f_{Gi}(t)$$

and if the quadratic criterion (2) is selected, then the term Φ vanishes altogether as

$$u = f_n^2(t) + 2f_n(t) \sum_{i=1}^{m} f_i(t)\,\Delta x_i + \sum_{i,\,j=1}^{m} f_{ij}(t)\,\Delta x_i \Delta x_j, \tag{5}$$

where in this case $f_i(t) = f_{Gi}(t)$, $f_{ij}(t) = f_{Gi}(t)\,f_{Gj}(t)$.

The deviations Δx_i can be represented as consisting of two parts: the search oscillations δx_i and the deviations Δx_i^* of the outputs from the integrations

$$\Delta x_i = \Delta x_i^* + \delta x_i.$$

In accordance with the latter, the expression (4) assumes the form

$$u = F(f_n) + 2f_n(t) \sum_{i=1}^{m} f_i(t)\,\Delta x_i^* + 2f_n(t) \sum_{i=1}^{m} f_i(t)\,\delta x_i + \sum_{i,\,j=1}^{m} f_{ij}(t)\,\Delta x_i^* \Delta x_j^* +$$
$$+ 2 \sum_{i,\,j=1}^{m} f_{ij}(t)\,\Delta x_i^* \delta x_j + \sum_{i,\,j=1}^{m} f_{ij}(t)\,\delta x_i \delta x_j + \Phi. \tag{6}$$

The output from the ith amplifier (Fig. 2) is equal to $u\delta x_i$. Having passed through the filter whose transfer function is $W_{fi}(D)$ and through an integrator, the signal $u\delta x_i$ represents the output from an integrator which, in accordance with out notation, is $x_{si} - \Delta x_i^*$. Thus

$$x_{si} - \Delta x_i^* = \frac{W_{fi}(D)}{D}(u\delta x_i)$$

or

$$D\Delta x_i^* = -W_{fi}(D)(u\delta x_i) + Dx_{ci}.$$

Using the expression (6), we obtain a complete system of equations for extremal reception

$$D\Delta x_i^* = -W_{fi}(D)\left\{2 \sum_{\nu,\,j=1}^{m} f_{\nu j}(t)\,\Delta x_\nu^* \delta x_j\,\delta x_i + 2f_n(t) \sum_{\nu=1}^{m} f_\nu(t)\,\delta x_\nu \delta x_i + \right.$$
$$+ 2f_n(t) \sum_{\nu=1}^{m} f_\nu(t)\,\Delta x_\nu^* \delta x_i + \sum_{\nu,\,j=1}^{m} f_{\nu j}(t)\,\Delta x_\nu^* \Delta x_j^* \delta x_i +$$
$$\left. + \sum_{\nu,\,j=1}^{m} f_{\nu j}(t)\,\delta x_\nu \delta x_j \delta x_i + F[f_n(t)]\,\delta x_i + \Phi \delta x_i \right\} + Dx_{si} \quad (i = 1, 2, \ldots, m). \tag{7}$$

These equations are nonlinear in the general case, the coefficients being ordinary ($f_\nu(t)$, $f_{\nu j}(t)$) as well as random ($f_n(t)$) functions of time. In the general case, it is not possible to integrate these equations as was not possible with other extremal regulations [2, 3]. Nonetheless, under some conditions, justifiable from the practical point of view, it is possible to obtain an approximate solution in a comparatively simple way.

Quasistationary Extremal Signal Reception

A behavior of considered systems, such that their parameters x_i, $\Delta x_i^{\bullet}$ vary slowly compared with $f_i(t)$, $f_{ij}(t)$, δx_i, shall be called quasistationary or, in other words, we shall say that the spectrum of the functions $\Delta x_i^{\bullet}(t)$ are sufficiently narrow.

The lower bound of the admissible spectrum width of the functions $\Delta x_i^{\bullet}$ is, of course, determined by the spectrum width of the signal parameter x_{si}, i.e., by the spectrum width of modulation. The latter, in turn, depends on the quantity of information transmitted in a unit of time. In a quasistationary state, the spectrum width or the spectrum "lines" of the signal useful component are known to be less than that of "carrier" frequencies or search frequencies.

In a quasistationary state, the passband of the filters W_{fi} is made to be narrow compared with the signal or search frequencies.

Quasistationarity of the behavior is the first assumption that we have made in order to be able to analyze the process.

The second assumption, closely related to the first one, is as follows. We represent the time functions entering the equations (7) in the form

$$\begin{aligned}
f_{\nu j}(t)\,\delta x_j \delta x_i &= \overline{f_{\nu j}(t)}\,\overline{\delta x_j \delta x_i} + \xi_{\nu j},\\
f_n(t) f_\nu(t)\,\delta x_\nu \delta x_i &= \overline{f_n(t) f_\nu(t)}\,\overline{\delta x_\nu \delta x_i} + \eta_{\nu i},\\
f_n(t) f_\nu(t)\,\delta x_i &= \overline{f_n(t) f_\nu(t)}\,\overline{\delta x_i} + \zeta_{\nu i},\\
f_{\nu j}(t)\,\delta x_i &= \overline{f_{\nu j}(t)}\,\overline{\delta x_i} + e_{\nu j},\\
f_{\nu j}(t)\,\delta x_\nu \delta x_j \delta x_i &= \overline{f_{\nu j}(t)}\,\overline{\delta x_\nu \delta x_j \delta x_i} + \tau_{\nu j i}.
\end{aligned} \tag{8}$$

Here the bar denotes the time average (constant components) of the respective quantities.

This condition implies that the spectra of the "rapidly varying" quantities ξ, η, ζ, ϵ, τ are outside the transmission bands of the filters W_{fi}, and the components are completely suppressed by those filters.

As far as the functions ξ, ϵ, and τ only are concerned, this condition can always ben ensured for a quasistationary state if an appropriate choice of the frequencies of search vibrations is made.

But with regard to the functions η and ζ in which the interference $f_n(t)$ is present, the above condition can be ensured when the interference spectrum is found within the spectrum domain of the carriers of the useful signal $f_{ij}(t)$, this being, however, the most important case. In other instances, one can obviously reduce it to the latter by introducing a sufficiently narrow-band filter at the system input.

We shall also assume that the search oscillations take place about zero and that they are orthogonal

$$\overline{\delta x_i} = 0, \quad \overline{\delta x_\nu \delta x_j \delta x_i} = 0, \quad \overline{\delta x_\nu \delta x_i} = \begin{cases} \delta x_\nu^2 & \text{for } \nu = i, \\ 0 & \text{for } \nu \neq i. \end{cases}$$

It is beyond the scope of this work to analyze fully the compatibility of these assumptions. For a number of specific instances, these assumptions have proved compatible (see the example cited below).

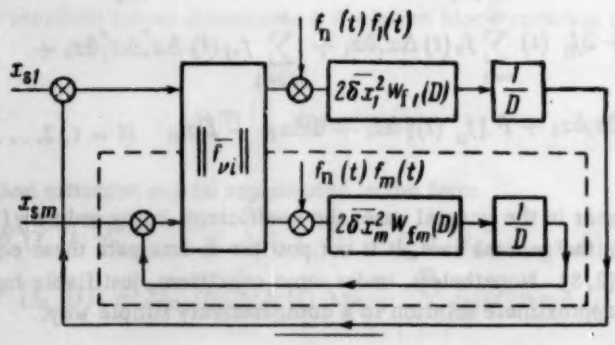


Fig. 4.

If our assumptions are valid, the equations (7) are transformed into

$$D\Delta x_i^{\cdot} + 2\overline{\delta x_i^2} W_{fi}(D) \sum_{\nu=1}^{m} \overline{f_{\nu i}(t)}\, \Delta x_{\nu}^{\cdot} =$$

$$= Dx_{si} - W_{fi}(D)\,[2f_n(t) f_i(t)\, \overline{\delta x_i^2} + F[f_n(t)]\,\delta x_i + \Phi\delta x_i]. \tag{9}$$

The spectrum of the quantity $F[f_n(t)]\delta x_i$ lies, in view of our assumptions with regard to $f_n(t)$, outside the transmission band of the filter N_{fi}.

The quantity $\Phi\delta x_i$ can certainly be neglected for small deviations Δx_i, in view of its higher order of smallness.

Therefore, in accordance with our assumptions, the equations for a quasistationary extremal reception degenerate into the following system of linear equations with constant coefficients:

$$D\Delta x_i^{\cdot} + 2\overline{\delta x_i^2} W_{fi}(D) \sum_{\nu=1}^{m} \overline{f_{\nu i}(t)}\, \Delta x_{\nu}^{\cdot} = Dx_{si} - 2\overline{\delta x_i^2} W_{fi}(D)\,(f_n f_i)$$

$$(i = 1, 2, \ldots, m). \tag{10}$$

The block-diagram corresponding to these equations is shown in Fig. 4, where $\|\overline{f_{\nu i}}\|$ is the matrix of coefficients.

It is apparent from the equations (10) and from the block-diagram that the processes of extremal adjusting of the generator are, under our assumptions, analogous to the processes in a linear multidimensional follow-up system whose input quantities are the parameters of the useful component of the input signal x_{si}, and whose outputs are the parameters of the generator signal x_i.

The role played by interference is taken here by the functions $f_n(t)$ $f_i(t)$. This enables one to perform the analysis and the synthesis of extremal reception in accordance with out assumptions on the same lines as the well-known methods for stationary linear systems.

The stability of the extremal reception is, in view of (10), determined from the characteristic equation

$$\begin{vmatrix} \dfrac{\lambda}{2\overline{\delta x_1^2} W_{f1}(\lambda)} + \overline{f}_{11} & \cdots & \overline{f}_{m1} \\ \cdots\cdots\cdots\cdots\cdots\cdots & & \\ \overline{f}_{1m} & \cdots & \dfrac{1}{2\overline{\delta x_m^2} W_{fm}(\lambda)} + \overline{f}_{mm} \end{vmatrix} = 0. \tag{11}$$

We shall say that the "carrier functions" $f_i(t)$ are orthogonal if the conditions are satisfied

$$\overline{f_{\nu i}(t)} = \begin{cases} \overline{f}_{\nu i} & \text{for} \quad i = \nu, \\ 0 & \text{for} \quad i \neq \nu. \end{cases} \tag{12}$$

When the carrier functions are orthogonal, the system (10) of equations breaks up into m independent equations

$$D\Delta x_i^{\cdot} + 2\overline{\delta x_i^2}\,\overline{f}_{ii} W_{fi}(D)\, \Delta x_i^{\cdot} = Dx_{si} - 2\overline{\delta x_i^2} W_{fi}(D)\,[f_n(t) f_i(t)]$$

$$(i = 1, 2, \ldots, m) \tag{13}$$

and the equivalent multidimensional follow-up system (Fig. 4) also breaks up into m independent channels.

The synthesis of the investigated extremal reception system consists in finding transfer functions of filters such that they ensure the lowest level of errors Δx_i for the known characteristics of the random function $f_n(t)$ and of ran-

dom or ordinary functions $x_{si}(t)$. It is customary to formulate the synthesis in this way and it is not necessary to dwell on it here.

As in other extremal regulations problems, these can be investigated in a sequence of approximations (of increasing complexity), starting from the simplest case of a quasistationary state [equations (10)] to the general case of forced self-adjustment (equations (7)]. Such an attempt was made in the paper [6] for one specific case of extremal regulation. The development, however, of some more powerful methods of analytic investigation of extremal regulation processes is a different problem.

It would be useful to clarify the important question of noise stability of quasistationary extremal reception compared with other well known types of reception.

Let the useful component of the received signal be of the form

$$f_s = \sum_{v=1}^{m} x_{sv} f_v(t),$$

where the carrier functions $f_v(t)$ are known without error.

One is able to apply successfully to such signals a correlation mode of reception or a synchronous detection method known as a "noisestable" method. In order to have such a reception, the signal f_s+f_n is multiplied by $f_i(t)$, and afterwards a filter is applied. Thus, in case of correlation mode of reception, one has

$$x_i = W_i(D)\left[\sum_{v=1}^{m} x_{sv} f_v(t) f_i(t) + f_n(t) f_i(t)\right].$$

As previously, transmission band of the low frequency filters W_i can be adopted such that they will suppress completely the variable components π_{vi} of the functions

$$f_v(t) f_i(t) = \overline{f_v(t) f_i(t)} + \pi_{vi}.$$

Thus

$$x_i = W_i(D)\left[\sum_{v=1}^{m} \overline{f_{vi}} x_{sv} + f_n(t) f_i(t)\right],$$

$$\overline{f_{vi}} = \overline{f_v(t) f_i(t)} \quad (i = 1, 2, \ldots, m). \tag{14}$$

When comparing these equation with those of quasistationary extremal reception in (10), it is easy to establish that with a specified relation between the transfer functions $W_{fi}(D)$ and $W_i(D)$, the two systems of equations are essentially the same.

Thus, if the carrier functions are orthogonal [as in (12)], in order that the equations (10) and (14) be identical it is sufficient that the condition

$$W_i(D) = \frac{2\overline{\delta x_i^2} W_{fi}(D)}{1 + 2\overline{\delta x_i^2}\,\overline{f_{ii}} W_{fi}(D)}, \tag{15}$$

be satisfied, that is, that it is enough that the transfer functions of the filters correspond to the operators of the closed and open loop single-circuit systems.

Hence, the conclusion that potential noise-stability of an extremal reception is equivalent to that of a correlation reception. The correlation reception can, however, take place only if the carrier functions $f_i(t)$ are known accurately, or more precisely, when these functions are implemented in the receiver.

An extremal receiver operating as a self-adjusting system does not require a preliminary reproduction of the carrier functions. This is a special feature of extremal reception and constitutes an advantage over correlation reception. An example will be given in order to illustrate this special feature.

Example. Let the useful component of the received signal take the form

$$f_s = A \sin(\omega t + \varphi).$$

We discuss two cases.

1. The carrier frequency ω is fixed and known exactly. The information is transmitted in the form of amplitude A and phase φ subjected to small variations not known beforehand.

We put the useful component of the received signal in the form $f_s = A\sin(\omega t+\varphi) = x_{s1}\sin\omega t + x_{s2}\cos\omega t$, with $x_{s1} = A\cos\varphi$, $x_{s2} = A\sin\varphi$ being the transformed parameters of the received signal, and $f_1(t) = \sin\omega t$ and $f_2 t = \cos\omega t$ the carrier functions.

The formula representing the self-adjusting generator of the extremal receiver shall be taken as

$$f_G = x_1 \sin\omega t + x_2 \cos\omega t,$$

where x_1 and x_2 are adjustment parameters.

With the quadratic deviations criterion (the block 1 in Fig. 2 now is a squaring device) the function u becomes

$$u = (f_s + f_n - f_G)^2 = (\Delta x_1 \sin\omega t + \Delta x_2 \cos\omega t + f_n)^2.$$

The search oscillations are considered harmonic with two different frequencies ω_1 and ω_2:

$$\Delta x_1 = \Delta x_1^* + \delta_m \sin\omega_1 t, \quad \Delta x_2 = \Delta x_2^* + \delta_m \sin\omega_2 t.$$

Then the equations (7) of extremal adjustment assume the form

$$\begin{aligned} D\Delta x_i^* = -W_{fi}(D)\,[&\delta_m^2 (2\Delta x_1^* \sin^2\omega t \sin\omega_1 t \sin\omega_i t + \Delta x_1^* \sin 2\omega t \sin\omega_2 t \sin\omega_i t + \\ &+ \Delta x_2^* \sin 2\omega t \sin\omega_1 t \sin\omega_i t + 2\Delta x_2^* \cos^2\omega t \sin\omega_2 t \sin\omega_i t) + \\ &+ 2f_m\delta_m (\sin\omega t \sin\omega_1 t \sin\omega_i t + \cos\omega t \sin\omega_2 t \sin\omega_i t) + 2f_m\delta_m (\Delta x_1^* \sin\omega t \sin\omega_i t + \\ &+ \Delta x_2^* \cos\omega t \sin\omega_i t) + \delta_m (\Delta x_1^{*2} \sin^2\omega t \sin\omega_i t + \Delta x_1^* \Delta x_2^* \sin 2\omega t \sin\omega_i t + \\ &+ \Delta x_2^{*2} \cos^2\omega t \sin\omega_i t) + \delta_m^3 (\sin^2\omega t \sin^2\omega_1 t \sin\omega_i t + \sin 2\omega t \sin\omega_1 t \cos\omega_2 t \sin\omega_i t) + \\ &+ f_n^2\,\delta_m \sin\omega_1 t] + Dx_{si} \quad (i = 1,\ 2). \end{aligned}$$

If the frequencies ω_1, ω_2 and ω are selected such that the combination of frequencies

$\omega_1 \pm \omega_2, \omega \pm \omega_1,\ \omega \pm \omega_2,\ 2\omega \pm \omega_1,\ 2\omega \pm \omega_2,\ \omega \pm \omega_1 \pm \omega_2,\ \omega \pm 2\omega_1,\ \omega \pm 2\omega_2,\ 2\omega \pm \omega_1 \pm \omega_2, 2\omega \pm 2\omega_1,\ 2\omega \pm 2\omega_2,$

as well as ω_1 and ω_2 are all outside the passband of the low-frequency filters W_{f1} and W_{f2}, then in a quasistationary state, all the components in the square bracket are very nearly suppressed except

and the equations become $\frac{1}{2}\Delta x_i^* \delta_m^2$, $f_n \delta_m^2 \sin\omega t$ (for $i = 1$), $f_n\,\delta_m^2 \cos\omega t$ (for $i = 2$), $f_n^2\,\delta_m \sin\omega_i t$,

$$D\Delta x_1^* + \frac{1}{2}\delta_m^2 W_{f1}(D)\,\Delta x_1^* = Dx_{s1} - W_{f1}(\delta_m^2 f_n \sin\omega t + \delta_m f_n^2 \sin\omega_1 t),$$

$$D\Delta x_2^* + \frac{1}{2}\delta_m^2 W_{f2}(D)\,\Delta x_2^* = Dx_{s2} - W_{f2}(\delta_m^2 f_n \cos\omega t + \delta_m f_n^2 \sin\omega_2 t), \qquad (16)$$

which is in full agreement with the equations (9) (the carrier functions satisfy here the orthogonality conditions (12), and $\Phi \equiv 0$).

If the spectrum of the interference is close to that of the useful signal, then the last components in the right-hand sides of (16) simply vanish and the equations of the quasistationary extremal reception assume the form

$$x_1 = W_1(D)\left(\frac{1}{2}x_{s1} - f_n \sin \omega t\right), \quad x_2 = W_2(D)\left(\frac{1}{2}x_{s2} - f_n \cos \omega t\right).$$

Exactly the same results can be obtained for correlation reception, if the received signal $x_{s1}\sin\omega t + x_{s2}\cos\omega t + f_n$ is multiplied by the carrier functions $\sin\omega t$ and $\cos\omega t$, and then sent through the narrow-band filters W_1 and W_2.

Because the correlation reception is more straightforward, it would be futile to use extremal reception. The extremum reception in a slightly different form should, however, be used when the carrier frequency is variable or unknown, and the correlation reception would not now be applicable.

2. Carrier frequency ω variable and not known precisely. Let the useful component of the received signal be as before, that is $f_s = A\sin(\omega t+\varphi) = A\sin\varphi_s$, $\varphi_s = \omega t+\varphi$ with the frequency unknown and in addition subjected to some random variations.

We take the frequency ω_G, or more precisely, the current phase $\varphi_G = \omega_G t+\varphi_0$, and also the amplitude A_G: $f_G = A_G\sin\varphi_G$ as adjustment parameters of the internal generator; this is ensured by an appropriate construction of the generator different from the one in the previous case. We assume that during the automatic adjustment the amplitude difference $\Delta A = A - A_G$, and also the difference of the current phases $\Delta\varphi = \varphi_s - \varphi_G$ are small. Then

$$f_s - f_G = A\sin\varphi_s - A_G\sin\varphi_G = \Delta A\sin\varphi_s + A\cos\varphi_s\,\Delta\varphi.$$

By denoting $\Delta x_1 = \Delta A$ and $\Delta x_2 = A\Delta\varphi$ and with other conditions being the same (harmonic search oscillations of ΔA and of $A\Delta\varphi$, quadratic criterion), it becomes obvious that the dynamics of this process is identical with the one we have just considered.

In particular, one can write down the equations analogous to the ones in (16):

$$\left[D + \frac{1}{2}\delta_m^2 W_{f1}(D)\right]\Delta A = DA - \delta_m^2 W_{f1}(D)(f_n \sin\omega t),$$

$$\left[D + \frac{1}{2}\delta_m^2 W_{f2}(D)\right]A\Delta\varphi = D(A\varphi_s) - \delta_m^2 W_{f2}(D)(f_n \cos\omega t),$$

where $\varphi_s = \omega t$.

It can be seen from these equations that when the transfer function $W_{f2}(D)$ has no poles at zero, that is when the corresponding filter contains no integrators in its single-circuit scheme, the system of extremal reception is static with regard to the change of frequency $\omega \approx D\varphi_s$ of the received signal. In order that it be astatic in frequency, it is necessary that the filters contain an integrator. As regards the change in amplitude A, the system is astatic when even the transfer function $W_{f1}(D)$ has no pole at zero.

In conclusion, it should be said that the signal reception is also solvable in the case of no a priori complete information on carrier frequencies, if appropriate follow-up systems are used.

Compared with the latter, the investigated method is certainly more flexible and versatile which, generally speaking, is inherent in self-adjusting systems.

LITERATURE CITED

1. A. A. Krasovskii, Dynamic continuous system of extremal regulation, based on the gradient method, Izv. AN SSSR, OTN. Énergetika i avtomatika No. 3 (1959).
2. A. A. Krasovskii, Statistical dynamic continuous extremal control system, Izv. AN SSSR, OTN. Énergetika i avtomatika No. 3 (1960).
3. A. A. Krasovskii, Statistical dynamic system with a proportional extremal self-adjusting series correcting device, Izv. AN SSSR, OTN. Énergetika i avtomatika No. 4 (1960).
4. A. A. Fel'dbaum, Calculating Devices in Automatic Systems [in Russian] (Fizmatgiz, 1959).
5. R. I. Stakhovskii, A multichannel automatic optimizer for solving variational problems, Avtomatika i telemekhanika 20, 1472 (1959).
6. V. S. Baranova and A. A. Pervozvanskii, Parametric phenomena in the simplest continuous extremal control systems, Avtomatika i telemekhanika 21, 1250 (1960).

A METHOD OF PLOTTING AMPLITUDE AND PHASE FREQUENCY CHARACTERISTICS WITH THE HELP OF A GENERALIZED FAMILY OF LOGARITHMIC CHARACTERISTICS

D. E. Polonnikov

(Moscow)

Translated from Avtomatika i Telemekhanika, Vol. 22, No. 6,
pp. 739-747, June, 1961
Original article submitted December 22, 1960.

A method is presented for plotting amplitude and phase frequency characteristics of electric networks for which the complex transfer coefficients can be represented in a canonical form. Tables for 28 circuits and the formulas which connect the circuit parameters with the parameters of a generalized family of logarithmic characteristics are given.

Amplitude and phase frequency characteristics* are the most important characteristics establishing the basic properties of electronic amplifiers. Knowledge of AFC and PFC are especially important for measuring and evaluating (operating amplifiers) in which a large amount of negative feedback is used so that the problem of ensuring stability and a broad passband becomes complicated. Determining AFC and PFC takes much time during the design of such amplifiers because their transfer functions are generally of a high order, and a calculation must be repeated several times before a desired shape of the characteristics is successfully attained.

It is much easier to use the method of logarithmic frequency characteristics [1 to 4] which permits the AFC and PFC of elementary sections to be constructed with the help of a set of prepared model curves, from which, in turn, the characteristics of an entire network are obtained by a simple summation. However, the method of logarithmic frequency characteristics as developed, was designed chiefly for the analysis of automatic regulating systems, and for the calculation of the elementary sections which are most prevalent in such systems. In electronic amplifiers a diversity of types of elementary sections are encountered for which families of characteristics or model curves are required. In addition, the association between circuit parameters and section parameters often proves to be so complicated that it is difficult to trace the direct dependence of the shapes of the AFC and PFC on the circuit parameters.

Presented below is a method of plotting AFC and PFC which permits a substantial simplification of the computing process so that the AFC and PFC of almost all the simplest networks encountered in amplifiers is reduced to only two generalized families of logarithmic characteristics. Moreover, each characteristic is determined by one parameter α which characterizes the attenuation of the transfer coefficient of the section for a frequency variation from 0 to ∞. The parameter α is physically descriptive, and is especially convenient for the synthesis of correcting networks.

1. Generalized Families of Logarithmic AFC and PFC

In Tables 1 to 4 there are given the basic circuits of the elementary RC and RL networks most prevalent in electronic amplifiers. All the networks given are related to a class of networks whose transfer coefficient modulus changes monotonically as a function of frequency, but not faster than 20 db/decade. The circuits of Tables 1 and 2 have a negative attenuation (the transfer coefficient increases with frequency) and have a leading phase angle. The circuits of Tables 3 and 4 have a positive attenuation and a lagging phase angle.

It can be shown that the complex transfer coefficient of any circuit of the first group ($\alpha > 1$, Tables 1 and 2) can be reduced to the following canonical form**

*Hereafter for brevity we will designate the characteristics referred to by AFC and PFC respectively.

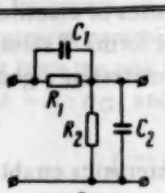
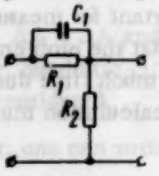
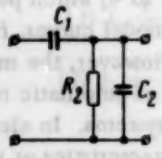
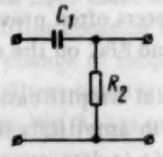
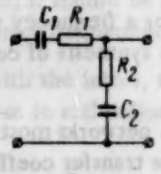
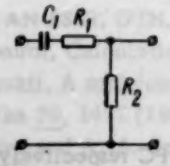
**It can be easily shown for each circuit that by substituting in (1) the values of X, α, and T from the tables given, we will get an expression for $K(j\omega)$ in the general form as written.

$$K(jX) = K(\infty)\,\frac{\frac{1}{\alpha-1} + jX}{\frac{\alpha}{\alpha-1} + jX}, \tag{1}$$

where $X = \omega T$ is a dimensionless variable; ω is the circular frequency; T is the equivalent time constant which is given in terms of circuit parameters in Tables 1 to 4; α is a parameter characterizing the change of $|K(jX)|$ for a change of ω from 0 to ∞,

$$\alpha = \left|\frac{K(\infty)}{K(0)}\right|. \tag{2}$$

TABLE 1

Circuits	$K(0)$	$K(\infty)$	α	T
	$\frac{R_2}{R_1+R_2}$	$\frac{C_1}{C_1+C_2}$	$\frac{C_1(R_1+R_2)}{R_2(C_1+C_2)}$	$\frac{R_1C_1}{\alpha-1} = R_1C_1\frac{R_2(C_1+C_2)}{R_1C_1-R_2C_2}$
	$\frac{R_2}{R_1+R_2}$	1	$1+\frac{R_1}{R_2}$	C_2R_2
	0	$\frac{C_1}{C_1+C_2}$	∞	$R_2(C_1+C_2)$
	0	1	∞	C_1R_2
	$\frac{C_1}{C_1+C_2}$	$\frac{R_2}{R_1+R_2}$	$\frac{R_2(C_1+C_2)}{C_1(R_1+R_2)}$	$\frac{C_2R_2}{\alpha-1} = C_2R_2\frac{C_1(R_1+R_2)}{C_2R_2-C_1R_1}$
	$\frac{C_1}{C_1+C_2}$	1	$1+\frac{C_2}{C_1}$	C_1R_2
	0	$\frac{R_2}{R_1+R_2}$	∞	C_1R_2

From (1) we obtain

$$|K(jX)| = K(\infty)\sqrt{\frac{\left(\frac{1}{\alpha-1}\right)^2 + X^2}{\left(\frac{\alpha}{\alpha-1}\right)^2 + X^2}}, \tag{3}$$

$$\tan\varphi = \frac{X}{\frac{\alpha}{(\alpha-1)^2} + X^2}, \tag{4}$$

where φ is the phase of K(jX).

TABLE 2

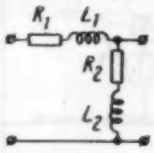
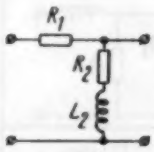
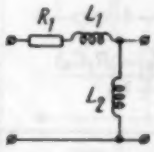
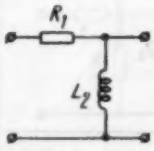
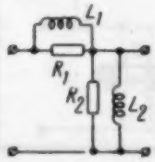
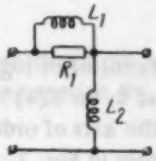
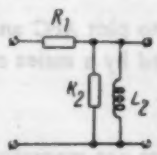
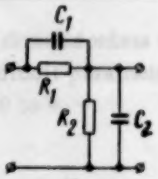
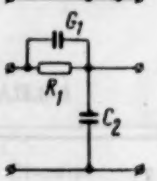
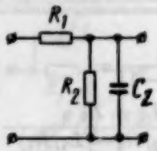
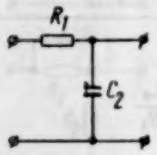
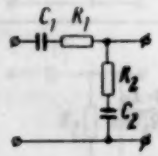
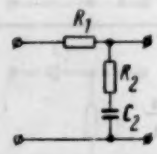
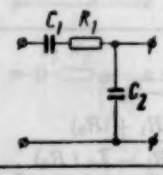
Circuits	$K(0)$	$K(\infty)$	α	T
	$\frac{R_2}{R_1+R_2}$	$\frac{L_2}{L_1+L_2}$	$\frac{L_2(R_1+R_2)}{R_2(L_1+L_2)}$	$\frac{\frac{L_2}{R_2(\alpha-1)}}{=\frac{L_2}{R_2}\frac{L_1+L_2}{R_1(L_2/R_2-L_1/R_1)}}$
	$\frac{R_2}{R_1+R_2}$	1	$1+\frac{R_1}{R_2}$	$\frac{L_2}{R_1}$
	0	$\frac{L_2}{L_1+L_2}$	∞	$\frac{L_1+L_2}{R_1}$
	0	1	∞	$\frac{L_2}{R_1}$
	$\frac{L_2}{L_1+L_2}$	$\frac{R_2}{R_1+R_2}$	$\frac{R_2(L_1+L_2)}{L_2(R_1+R_2)}$	$\frac{\frac{L_1}{R_1(\alpha-1)}}{=\frac{L_1L_2(R_1+R_2)}{R_1^2R_2(L_1/R_1-L_2/R_2)}}$
	$\frac{L_2}{L_1+L_2}$	1	$1+\frac{L_1}{L_2}$	$\frac{L_2}{R_1}$
	0	$\frac{R_2}{R_1+R_2}$	∞	$\frac{L_2}{R_2}\left(1+\frac{R_2}{R_1}\right)$

TABLE 3

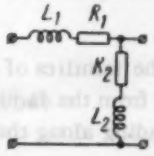
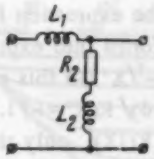
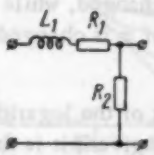
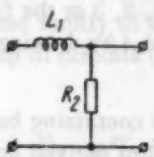
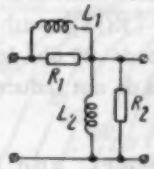
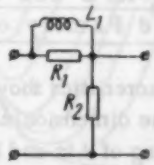
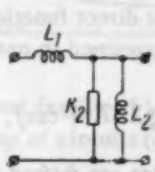
Circuits	$K(0)$	$K(\infty)$	α	T
	$\frac{R_2}{R_1+R_2}$	$\frac{C_1}{C_1+C_2}$	$\frac{C_1(R_1+R_2)}{R_2(C_1+C_2)}$	$R_1C_1\left(\frac{1}{\alpha}-1\right)=$ $=\frac{R_1(R_2C_2-R_1C_1)}{R_1+R_2}$
	1	$\frac{C_1}{C_1+C_2}$	$\frac{C_1}{C_1+C_2}$	C_2R_1
	$\frac{R_2}{R_1+R_2}$	0	0	$C_2R_1\frac{R_2}{R_1+R_2}$
	1	0	0	C_2R_1
	$\frac{C_1}{C_1+C_2}$	$\frac{R_2}{R_1+R_2}$	$\frac{R_2(C_1+C_2)}{C_1(R_1+R_2)}$	$R_2C_2\left(\frac{1}{\alpha}-1\right)=$ $=\frac{C_2(R_1C_1-R_2C_2)}{C_1+C_2}$
	1	$\frac{R_2}{R_1+R_2}$	$\frac{R_2}{R_1+R_2}$	C_2R_1
	$\frac{C_1}{C_1+C_2}$	0	0	$C_2R_1\frac{C_1}{C_1+C_2}$

By changing the value of the parameter α, it is possible to construct generalized families of logarithmic AFC and PFC for the circuits of the first group (Tables 1 and 2) which are shown in Figs. 1 and 2 for $K(\infty) = 1$. Inasmuch as the quantity $\lg K(\infty)$ is a component on a logarithmic scale, the initial reading along the axis of ordinates is shifted by a quantity $\lg K(\infty)$ which can be taken into account with the help of the $K(\infty)$ scale shown in Fig. 1.

As will be shown below, the generalized families of characteristics make it easy to plot AFC and PFC for any circuit of Tables 1 and 2, as well as for any complicated network that can be represented by a series connection of such circuits which are separated by detecting elements* (for example, electron tubes).

*In a number of cases it is permissible to break circuits down into sections when they are not separated by detecting elements provided that the output impedance of a preceding section is at all frequencies much smaller than the input impedance of the following section.

TABLE 4

Circuits	$K(0)$	$K(\infty)$	α	T
	$\frac{R_2}{R_1+R_2}$	$\frac{L_2}{L_1+L_2}$	$\frac{L_2(R_1+R_2)}{R_2(L_1+L_2)}$	$\frac{L_2}{R_2}\left(\frac{1}{\alpha}-1\right)=$ $=\frac{(L_1/R_1-L_2/R_2)R_1}{R_1+R_2}$
	1	$\frac{L_2}{L_1+L_2}$	$\frac{L_2}{L_1+L_2}$	$\frac{L_1}{R_2}$
	$\frac{R_2}{R_1+R_2}$	0	0	$\frac{L_1}{R_1+R_2}$
	1	0	0	$\frac{L_1}{R_2}$
	$\frac{L_2}{L_1+L_2}$	$\frac{R_2}{R_1+R_2}$	$\frac{R_2(L_1+L_2)}{L_2(R_1+R_2)}$	$\frac{L_1}{R_1}\left(\frac{1}{\alpha}-1\right)=$ $=\left(\frac{L_1}{R_1}-\frac{L_2}{R_2}\right)\frac{L_1}{L_1+L_2}$
	1	$\frac{R_2}{R_1+R_2}$	$\frac{R_2}{R_1+R_2}$	$\frac{L_1}{R_2}$
	$\frac{L_2}{L_1+L_2}$	0	0	$\frac{L_2}{R_2}\frac{L_1}{L_1+L_2}$

For the second group of circuits (Tables 3 and 4) having a positive attenuation ($\alpha<1$) it is convenient to use the following expression for the transfer coefficient

$$K(jX)=K(0)\frac{1+jX\frac{\alpha}{1-\alpha}}{1+jX\frac{1}{1-\alpha}}, \qquad (5)$$

whence

$$|K(jX)|=K(0)\sqrt{\frac{1+\left(\frac{\alpha X}{1-\alpha}\right)^2}{1+\left(\frac{X}{1-\alpha}\right)^2}}, \qquad (6)$$

$$\tan\varphi = -\frac{X}{1+\frac{\alpha}{(1-\alpha)^2}X^2}. \qquad (7)$$

Families of generalized logarithmic characteristics for the second group of circuits (Tables 3 and 4), plotted according to formulas (6) and (7), are shown in Figs. 3 and 4.

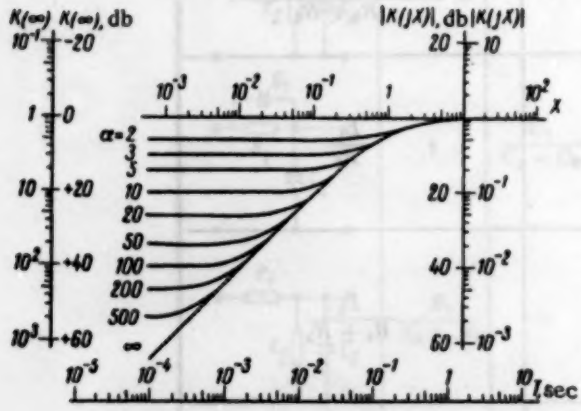


Fig. 1. Generalized family of logarithmic AFC of the first group of circuits ($\alpha > 1$, Tables 1, 2).

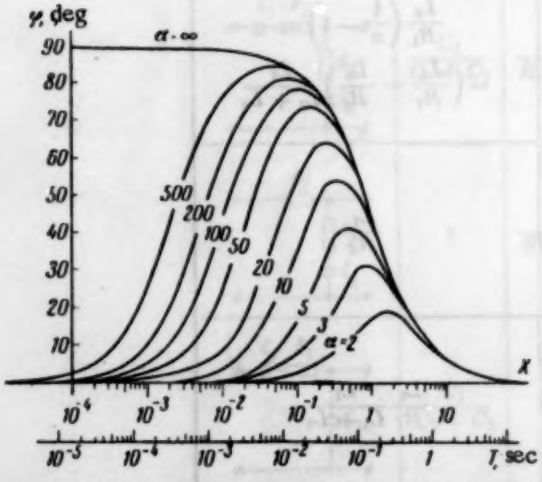


Fig. 2. Generalized family of logarithmic PFC of the first group of circuits ($\alpha > 1$, Tables 1, 2).

It is easily seen that the families of characteristics shown in Figs. 3 and 4 differ from the families of Figs. 1 and 2 only in the direction of reading along the coordinate axes.

It can be shown that the expression for the complex transfer coefficient (5) transforms into expression (1) upon replacing the variable X by $-1/X'$ [in this case $\alpha = K(\infty)/K(0)$ becomes equal to $K(X'=0)/K(X'=\infty)$ i.e. $1/\alpha'$]. This implies that for the family $|K(jX)|$ only the direction of reading along the X axis is changed, while for the family $\varphi(X)$, in addition, the sign of φ is changed because $\varphi(X)$ is an odd function.

Thus, for a useful plot of the logarithmic AFC and PFC of any of the 28 circuits given, it is sufficient to have either the families of Figs. 1, 2, 3 or the families of Figs. 1, 3, 4, i.e. to unite Figs. 2 and 4 by drawing additional scales in the opposite direction.

We note that networks containing both L and C cannot be calculated from the families derived because the maximum rate of attenuation of $|K(jX)|$ with a change in frequency for such circuits exceeds 20 db/decade, and their complex transfer coefficients do not reduce to the forms of Eqs. (1) or (5).

2. Method of Plotting AFC and PFC with the Help of a Generalized Family of Logarithmic Characteristics

The logarithmic characteristics shown in Figs. 1 to 4 are plotted as functions of the dimensionless variable $X = \omega t = 2\pi fT$. For practical plotting of Afc and PFC it is desirable to have the characteristics as direct functions of the circular frequency ω or frequency f expressed in cps. Inasmuch as

$$\lg X = \lg 2\pi T + \tan f, \qquad (8)$$

the logarithmic characteristics are shifted along the axis of abscissas by the quantity $\lg 2\pi T$, which is equivalent to a shift of the origin of coordinates by $-\lg 2\pi T$. Allowing for this, it is possible to plot beforehand a scale of time constants that show where the origin of coordinates should be located along the axis of abscissas for a given value of T. Such scales are laid off along the axes of T in Figs. 1 to 4.

There is presented below the sequence for constructing logarithmic AFC and PFC for any circuit given in Tables 1 to 4.

First, from the formulas given in a table, T, α, K(0) and $K(\infty)$ are determined, and two coordinate grids are prepared as shown in Fig. 5* (separately for AFC and PFC). Then the coordinate grid of AFC is superposed on the

*The scale of the coordinate grid must conform closely with that shown in Fig. 5, seeing that it agrees with the scale of the curves of Figs. 1 to 4. The length of the axes depends on the particular case. It is expedient to make them cover a considerably larger range of variation of f, K and φ than is shown in Fig. 5.

corresponding generalized family of AFC (Figs. 1 or 3) so that the axis of ordinates will stand on the mark of scale T corresponding to the time constant T* of the circuit, and the axis of abscissas is fixed at the mark corresponding to the given value of K(∞) or K(0) along the vertical scale. After this, the curve for the required value of α is plotted on tracing paper. In analogous fashion, the PFC is plotted, but for this case the axis of abscissas on the tracing paper must always coincide with the axis of abscissas on Figs. 2 or 4.

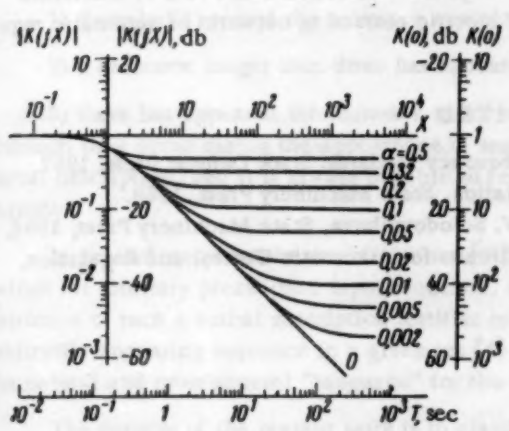


Fig. 3. Generalized family of logarithmic AFC of the second group of circuits (α < 1, Tables 3, 4).

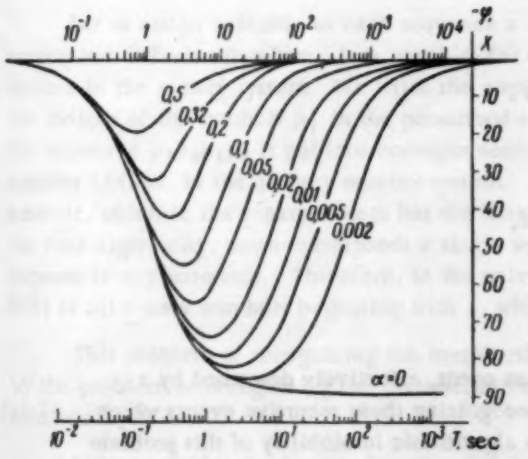


Fig. 4. Generalized family of logarithmic PFC of the second group of circuits (α< 1, Tables 3, 4).

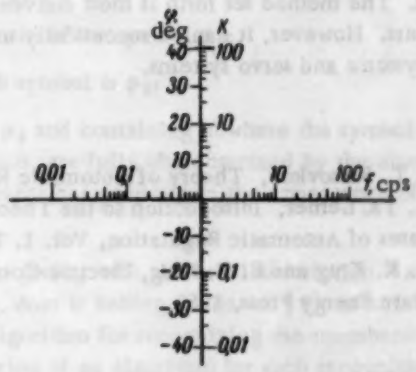


Fig. 5. Example of a coordinate grid for plotting AFC and PFC

After plotting the characteristics of all the series-connected sections, their simple sum is taken, and the AFC and PFC of the comlete network being considered is obtained.

The process of plotting AFC and PFC can be made even easier by preparing two sets of model curves.

The accuracy of the AFC and PFC plots is determined chiefly by the accuracy of plotting the characteristics of the generalized family (or model curves), and also by the errors which sometimes occur during the analysis of a circuit into elementary sections. As experience has shown, for a plot of logarithmic characteristics by this method, the total error can be held to a value not exceeding 0.5 db for the AFC and 2 to 3° for the PFC.

To the advantages of the present method besides high accuracy must be added its simplicity. For a comparatively large class of networks there are used only two families of characteristics, which significantly speeds up the process of plotting AFC and PFC. In addition, the use of the descriptive parameter α = K(∞)/K(0) significantly helps in the selection of parameters for the various sections during the synthesis of a complicated circuit for a given AFC and PFC.

The method presented was developed chiefly for the analysis and synthesis of AFC and PFC of electronic amplifiers for which the equivalent circuits can be represented in the form of a series connection of elementary sections as given in Tables 1 to 4. However, the advantages of the method enumerated above apply to the calculation of any other electrical networks which can be represented as a series connection of circuits of the class under consideration, especially to the calculation of correcting networks for systems of automatic regulation and servo systems.

*In case the required value of T lies outside the scale of Figs. 1 to 4, the T scale should be extended in the necessary direction with the adopted logarithmic scale.

SUMMARY

1. The method presented permits the reduction of calculations for amplitude and phase frequency characteristics of a whole class of elementary sections (28 circuits) to the selection and copying of the required characteristics from a family of generalized logarithmic characteristics which are calculated beforehand. Because of this, great rapidity of synthesis, while retaining accuracy, is ensured.

2. The method set forth is most convenient for the analysis and synthesis of the AFC and PFC of electronic amplifiers. However, it can be successfully used for the synthesis of electric correcting networks of automatic regulation systems and servo systems.

LITERATURE CITED

1. I. T. Kuzovkov, Theory of Automatic Regulation Based on Frequency Methods, State Defense Press, 1957.
2. A. Ya. Lemer, Introduction to the Theory of Automatic Regulation, State Machinery Press, 1958.
3. Bases of Automatic Regulation, Vol. 1, Theory, edited by V. V. Solodovnikova, State Machinery Press, 1954.
4. G. K. Krug and E. K. Krug, Electric Correcting Elements in Circuits for Automatic Control and Regulation, State Energy Press, 1959.

THE ALGORITHMIC INSOLUBILITY OF THE PROBLEM OF RECOGNIZING THE REPRESENTABILITY OF RECURSIVE EVENTS IN FINITE AUTOMATA

M. A. Aizerman, L. A. Gusev, L. I. Rozonoér,

I. M. Smirnova, and A. A. Tal'

(Moscow)

Translated from Avtomatika i Telemekhanika, Vol. 22, No. 6,
pp. 748-755, June, 1961
Original article submitted January 4, 1961.

The concept of recursive event, formalizing the concept of "an event, effectively described by a verbal description," is introduced. The problem is posed of recognizing those recursive events which may be represented on a finite automaton. A theorem on the algorithmic insolubility of this problem is proved. At the same time, the concept of representation of a recursive function in an automaton is introduced and a criterion for recognizing representable functions is given.

As this group of problems is being considered in "Avtomatika i Telemekhanika" for the first time, certain well-known concepts and facts are also explained.

1. Recursive Events

In the notation used in [1, 2], the symbols ρ_i from the alphabet $\{\rho\} = \{\rho_0, \rho_1, \ldots, \rho_{r-1}\}$ denote the input states of a finite automaton. By definition, the number of states is finite and equal to r. In the present article we consider only nonautonomous finite automata such that $r \geq 2$.

An arbitrary sequence composed of a finite number of symbols ρ_i, for example

$$\rho_5,$$
$$\rho_5\rho_1\rho_3,$$
$$\rho_5\rho_1\rho_5\rho_1\rho_5\rho_4$$
$$\text{etc.},$$

will be termed the input sequence. The number of symbols constituting the sequence will be termed its length.

Let in any way, in an infinite ensemble $\{E\}$ of all conceivable input sequences (each of which has arbitrary, but finite, length), a finite or infinite subset $\{S\}$ of input sequences be distinguished. Let now there appear in succession certain input sequences from $\{E\}$. We say that an event S has occurred if the sequence appearing belongs to $\{S\}$, and that the event has not occurred if the sequence appearing does not belong to $\{S\}$. If the set $\{S\}$ is finite, the event S can be given simply by counting all sequences which enter into $\{S\}$. If the set $\{S\}$ is infinite, this method of prescribing the event is excluded, and the set must be defined in some other way. Thus, for example, the events:

1) there has appeared the sequence containing in any arbitrary position the symbol ρ_5;

2) a sequence longer than three has appeared; in which the fourth symbol is ρ_2;

3) there has appeared the sequence terminating with the symbol ρ_8 and containing nowhere the symbol ρ_5 — although they occur during the appearance of sequences from infinite sets, are fully characterized by the above verbal description, and it is always possible to recognize whether the event has occurred in the appearance of any sequence.

In general, to prescribe an infinite set $\{S\}$ of sequences signifies giving such a "verbal description," as will permit for arbitrary preassigned input sequence, to answer the question, does it belong to the set $\{S\}$ or not ? The existence of such a verbal description testifies to the existence of an algorithm for recognizing the membership of an arbitrarily appearing sequence in a given set $\{S\}$. In this sense the giving of an algorithm for such recognition is the natural and most general "language" for the effective prescription of infinite sets $\{S\}$, i.e., for prescribing events.

The purpose of the present work is to clarify to what degree a finite automaton can realize such recognition. To obtain an answer to this question it is necessary to formalize the intuitive concepts introduced above of "verbal description" and "recognition algorithm." A formalization of the concept "algorithm," as is well known, is given by the theory of recursive functions and Church's thesis.

Let us assign uniquely to each sequence a certain arbitrary number, such that to different sequences there correspond different numbers. It is possible, for example, to use the following coding of sequences by numbers, written in the r-nary system. We write the number 1 and immediately after it from left to right we write in sequence the indices of the symbols ρ_i in the prescribed sequence, reading from right to left. Thus, for example, with $r = 5$, the sequence $\rho_2\rho_0\rho_1\rho_3$ is put into correspondence with the number 13102, while the sequence $\rho_2\rho_3\rho_0\rho_4\rho_1$, with the number 114032, in the quinary number system. In this coding system, to each sequence there corresponds a single number, which in the r-nary system has the first digit 1, and vice-versa, to each number, having in the r-nary system the first digit unity, there corresponds a single sequence. Numbers whose first digit is not equal to unity do not correspond to any sequence. Therefore, to the universal set $\{E\}$ of all conceivable sequences there corresponds the set $\Lambda(E)$ of all r-nary numbers beginning with 1, while to the set $\{S\}$, a certain subset $\Lambda(S)$ of these numbers.

This problem of recognizing the membership of a given sequence in the set of sequences $\{S\}$ is now substituted by the problem of recognizing the membership of the corresponding r-nary number (beginning with 1) in the set of numbers $\Lambda(S)$.

Let us consider the integer function $\psi_S(x)$ defined on the set of r-nary numbers beginning with 1, and taking the values 1 for all numbers in $\Lambda(S)$, and 0 for all remaining numbers. If there exist an algorithm for recognizing membership of the sequence in the set $\{S\}$, the values of the function $\psi_S(x)$ can be calculated for all x in $\Lambda(E)$. In this sense the function $\psi_S(x)$ is calculable.

In agreement with Church's thesis, asserting the partial recursiveness of each integer function, calculated for values of an argument from a certain set, the function $\psi_S(x)$ is partially recursive,* defined in $\Lambda(E)$.

*In the theory of algorithms an integer function is termed generally recursive (or frequently, simply "recursive"), if it can be obtained from certain more simply given functions (identity functions, constant functions, etc.) only by means of certain defined operations (superposition, primitive recursion, etc.). The function $\psi(x)$, defined in a certain subset of integers, is termed partially recursive if there exists a recursive function $\varphi(x)$, taking on the same values as $\psi(x)$ for values of the argument taken from this subset. Church's thesis declares the equivalence of the concepts: recursive function, calculable function, algorithm (cf. [3, 4]).

Let us construct the integer function $\Psi_S(x)$, already defined among all integers, the values of which coincide with $\Psi_S(x)$ for $x \in \Lambda(E)$, and is equal to 0 for all other x.

In regard to the existence of an elementary algorithm for testing the membership of the number x in the set $\Lambda(E)$ (writing x in the r-nary number system and determining the presence of the 1 in front), and the partial recursiveness of $\Psi_S(x)$, the function $\Psi_S(x)$ is recursive. From this it also follows that the function $\Psi_E(x)$, taking the value 1 for $x \in \Lambda(E)$ and 0 for the remaining x, is recursive.

A function given for all values of an integer argument, and taking the value 1 in a certain set $\{M\}$ of arguments and 0 outside it is termed the characteristic function for $\{M\}$ sets of numbers whose characteristic functions are recursive are termed recursive sets.

We term the event S recursive if the set $\Lambda(S)$ is recursive. The universal event E, consisting of the appearance of any arbitrary sequence of symbols from $\{E\}$, is recursive, since the function $\Psi_E(x)$ is recursive.

Above it was shown that in the case where the event S is such that there exists an algorithm for recognizing the fact that event has occurred (i.e., the membership of any given sequence in the set $\{S\}$), the characteristic function $\Psi_S(x)$ of the set $\Lambda(S)$ is recursive, i.e., $\Lambda(S)$ is a recursive set.

Thus (this is a direct consequence of Church's thesis!), each event for which a "recognition rule" can be stated, is recursive, and vice-versa.

The concept of recursive event is the broadest and most natural formalization of the intuitive concept "event which can be effectively described," "event for which a recognition rule can be given," etc.*

2. Representation of Events in an Automaton and the Problem of Recursive Event Representability

Les us consider a finite autoton A (cf. [1, 2]), described by the recursive relation

$$\varkappa^{p+1} = F(\varkappa^p, \rho^p), \quad p = 0, 1, 2, \ldots, \tag{1}$$

where F is a single-valued function, put into correspondence with the pair of symbols ρ_i (from the input alphabet $\{\rho\}$) and $\varkappa_j$ (from the alphabet of states $\{\varkappa\} \equiv \{\varkappa_0, \ldots, \varkappa_{q-1}\}$), a symbol from $\{\varkappa\}$. In the set $\{\varkappa\}$ we define a certain subset M. We shall assume that automaton (1) begins operating at the time p = 0 in the initial state $\varkappa(0) = \varkappa^0 \in \{\varkappa\}$, under the action of the sequence $\rho^0\rho^1\rho^2\ldots$, developing from cycle to cycle. In each given cycle p the sequence of input symbols appearing from the start of operation of the automation to the present cycle, has a finite length $p + 1$.

Let us assume now that an event S is given, in the sense in which this term was used in Sect. 1. It can happen that up to the time p the input symbol sequence belongs to the set $\{S\}$, i.e., that at this instant "the event has occurred." The event S is termed representable in a finite automaton if there exists such an automaton A, initial state $\varkappa^0$ and set of states $M \subset \{\varkappa\}$, such that the sequence of input symbols up to the time p belongs to $\{S\}$ if and only if the automaton is in one of the states belonging to M at the time $p + 1$.

The automaton A can have an output decoder

$$\lambda^p = \Phi(\varkappa^p),$$

which, for all symbols $\varkappa_j \in M$, puts out $\lambda^p = 1$, while for the remaining symbols, $-\lambda^p = 0$. Then the appearance of the symbol 1 at the output of this "automaton with decoder" indicates the occurrence S.

The problem naturally arises of determining the class of events representable in a finite automaton, and their discrimination from events which cannot be represented in any finite automaton.

* If the requirement of effectiveness of describing the event is not put, it is possible to prescribe the events in the still broader language of recursive-denumerable sets. In such a description, however, the algorithmic unsolvability already extends to the problem of recognizing whether an event has occurred with the appearance of some given sequence, and, a fortiori, and problem formulated below of recognizing the representability of events is algorithmicly unsolvable.

Kleene [5] has introduced the concept of regular event, and has proved that for an event to be representable in a finite automaton, it is necessary and sufficient that it be regular.

In the terminology of the present work,* the following definition can be given for regular event.

Three operations on sets of sequences are considered. Let there be given the sets of sequences $\{S\}$ and $\{Q\}$.

A. The union $W = S \cup Q$ is the set containing all sequences occurring in at least one of the sets S or Q.

B. The product W = S × Q is the set containing all sequences obtained by the adjoining to each sequence from S at the right one each of the sequences from Q.

C. The interation $W = S^*$ is the set containing all sequences obtained by adjoining at the right to each sequence from S any number of arbitrary sequences from S itself.

The set consisting of an arbitrary sequence of length 1 (a single symbol ρ_i), is termed an elementary set. By definition, elementary sets, and all sets which can be obtained from elementary sets by any finite number of the above three operations, are regular.

If the set $\{S\}$ is regular, then the event S corresponding to it is also termed regular.

It can be considered that an event is given directly in the language of regular sets. In this case, by virtue of Kleene's theorem, no question arises as to whether it can be represented by an automaton. However, an event at the input can be given in any other manner, for example, by a verbal description, giving the instruction for recognizing the occurrence of the event. To answer the question whether this event is representable in an automaton, it is necessary to clarify if it is regular. Thus, the problem arises of defining the regularity of recursive event. It is easily proven that any regular event is recursive. But the inverse statement is not true, since recursive irregular (i.e., nonrepresentable!) events are known.

The problem of present interest may be formulated as follows: does there exist an algorithm permitting, for, an arbitrary recursive event, an answer to the question of its regularity (i.e., its representability)? We term this problem the problem of recognizing the representability of recursive events.

3. The Algorithmic Unsolvability of the Problem of Recognizing the Representability of Recursive Events

In this section we shall be concerned with the phenomenon of algorithmic unsolvability. A problem is termed algorithmically unsolvable if the nonexistence of an algorithm solving it can be proven. By virtue of Church's thesis, the concept "algorithm" corresponds to the concept "recursive function." This signifies that in each concrete case all possible initial data and results of solving the problem are coded by integers, and the existence of an algorithm solving the problem corresponds to the existence of a recursive function, which takes on values coding the results of solving the problem when the argument takes on values coding the "initial data." Therefore, the problem is algorithmically unsolvable if such a recursive function does not exist.

Theorem 1. The problem of recognizing the representability of recursive events is algorithmically unsolvable.

* Kleene has operated with sets of input tables, but to each table there may correspond a sequence of input symbols. For Kleene these sequences have developed, as it were, "from right to left," beginning from the instant p

$$\ldots \rho(p-2),\ \rho(p-1),\ \rho(p),$$

in contradistinction to the present work, where the sequences have developed "from left to right"

$$\rho(0),\ \rho(1),\ \rho(2)\ \ldots$$

In Kleene's terminology this signifies, in particular, that we are considering primitive events, and the definition of regular operation changes somewhat (see below). A similar presentation is given in [6].

After the present authors gave a proof Theorem 1, it was found that the statement contained in this theorem was given earlier without proof in 5° of B. A. Trachtenbrot's work [7]; we shall, therefore, in what follows term this theorem Trachtenbrot's theorem.

To prove Trachtenbrot's theorem, we shall first formulate a more restricted problem than the problem of recognizing the representability of recursive events, and we shall prove that even this narrower problem is algorithmically unsolvable. This in itself will prove the theorem. Let there be given a recursive function $\rho(t)$, defined on a set of integers, and taking on values from the finite set of numbers $\{0, 1, \ldots, r-1\}$. We consider an automaton A with input alphabet $\{\rho_0, \ldots, \rho_{r-1}\}$. Of all possible input sequences we shall note especially the sequence

$$\rho_{\varphi(0)},$$
$$\rho_{\varphi(0)}\rho_{\varphi(1)},$$
$$\rho_{\varphi(0)}\rho_{\varphi(1)}\rho_{\varphi(2)}$$
etc.

where $\rho_{\varphi(i)}$ is a symbol from $\{\rho\}$ whose index coincides with the value of the recursive function $\varphi(t)$ at $t = i$. We consider the event S^{φ}, consisting in that up to the given instant there has appeared at the input to the automaton any one of the above-defined sequences. We say that the automaton represents the recursive function $\varphi(t)$ if it represents the event S^{φ}. In other words, the automaton represents the recursive function $\varphi(t)$ if at each instant of time p the state of the automaton $\varkappa(p)$ belongs to a given set M if and only if the indices of all input symbols appearing between $t = 0$ and $t = p$ are sequences of values of the function $\varphi(t)$.

Obviously there exist recursive functions certainly representable (for example, any periodic function, since for if the event S^{φ} is regular) and certainly nonrepresentable (for example, the function $\alpha(t)$, taking the value 1 for $t = n^2$ and 0 in the remaining cases). The problem therefore arises of recognizing the representability of recursive functions. It is easily seen that this problem is a particular case of our general problem of recognizing the representability of recursive functions.

The function $\varphi(t)$ will be termed "finally periodic" if there exists a pair of numbers τ and T, such that for arbitrary $t \geq \tau$, the condition

$$\varphi(t+T) = \varphi(t),$$

is satisfied, i.e., the value of the function for $t \geq \tau$ repeats periodically with period T.

Theorem 2. The function $\varphi(t)$ is representable in a finite automaton if and only if it is finally periodic.

We note that Theorem 2 permits constructing an unlimited number of examples of nonrepresentable events. What is essential for us here is that this theorem reduces the problem of recognizing the representability of recursive functions to the problem of recognizing whether a given recursive function is finally periodic. But since even this more restricted problem is algorithmically unsolvable,* the proof of Theorem 2 completes the proof of Trachtenbrot's theorem as well.

We now proceed to prove Theorem 2.

Proof of Theorem 2. We shall first prove that if the function $\varphi(t)$ is finally periodic, it is representable in a finite automaton. For this purpose we prove that the set $\{S^{\varphi}\}$ of sequences for whose appearances the event occurs, is regular.

Let to the period of the function $\varphi(t)$ for $t \geq \tau$ these correspond the sequence of input symbols $\rho_{k_1}, \rho_{k_2}, \ldots, \rho_{k_T}$. Then the set $\{S^{\varphi}\}$ can be represented in the form

$$S^{\varphi} = (\rho_{\varphi(0)} \times \rho_{\varphi(1)} \times \ldots \times \rho_{\varphi(\tau-1)}) \times (\rho_{k_1} \times \ldots \times \rho_{k_T})^{*} \times [\rho_{k_1} \cup (\rho_{k_1} \times \rho_{k_2}) \cup \ldots \cup (\rho_{k_1} \times \ldots \times \rho_{k_T})],$$

i.e., it can be expressed in terms of elementary sets of sequences, using the regular operations, Q.E.D.

We shall now prove the converse statement, i.e., that if the function $\varphi(t)$ is representable in a finite automaton, it is finally periodic. Let A be a finite automaton having q different states, representing the function $\varphi(t)$, with initial state $\varkappa_0$ a property of the state M.

* The proof of this fact is contained, for example, in Chapter 10 of [8], where it is carried out for the case where the function $\varphi(t)$ takes on two values; the extension of this proof to the general case is obvious.

Let us consider the input sequence of arbitary length $p > q$:

$$p_{\varphi(0)} p_{\varphi(1)} \cdots p_{\varphi(q)} \cdots p_{\varphi(p)} \tag{2}$$

and the sequence of automaton states corresponding to it

$$\varkappa^0 \varkappa(1) \varkappa(2) \ldots \varkappa(q+1) \ldots \varkappa(p+1), \tag{3}$$

satisfying the recurrent relationship (1).

Since automaton A has only q different states, in the sequence

$$\varkappa(1) \varkappa(2) \ldots \varkappa(q+1) \tag{4}$$

of length $q + 1$, at least two symbols will coincide, i.e., there exist numbers α, β ($\alpha < \beta$, $1 \leqslant \alpha < q+1$, $1 < \beta \leqslant q+1$), such that

$$\varkappa(\alpha) = \varkappa(\beta). \tag{5}$$

Let us consider the function

$$\widetilde{\varphi}(t) = \begin{cases} \varphi(t) & \text{for } 0 \leqslant t < \alpha, \\ \varphi(t+\beta-\alpha) & \text{for } \alpha \leqslant t < \infty, \end{cases} \tag{6}$$

whose values for $t = 0, 1, \ldots, \alpha-1$ are defined by the values $\varphi(0), \varphi(1), \ldots, \varphi(\alpha-1)$, , while for $t = \alpha$, $\alpha+1$, $\alpha+2, \ldots$, by the values $\varphi(\beta), \varphi(\beta+1), \varphi(\beta+2)$,..., i.e., the sequences $\widetilde{\varphi}(\alpha)\widetilde{\varphi}(\alpha+1) \ldots \widetilde{\varphi}(\alpha+s)$ and $\varphi(\beta)\varphi(\beta+1) \ldots \varphi(\beta+s)$ coincide for arbitrary s.

We shall now consider the sequence of input symbols

$$p_{\widetilde{\varphi}(0)} p_{\widetilde{\varphi}(1)} \cdots p_{\widetilde{\varphi}(\alpha-1)} p_{\widetilde{\varphi}(\alpha)} \cdots p_{\widetilde{\varphi}(\beta)} \cdots p_{\widetilde{\varphi}(p)} \tag{7}$$

and the sequence of states

$$\varkappa^0 \widetilde{\varkappa}(1) \ldots \widetilde{\varkappa}(\alpha-1) \widetilde{\varkappa}(\alpha) \ldots \widetilde{\varkappa}(p+1) \tag{8}$$

corresponding to it, by virtue of (1).

Since the first α symbols in sequences (2) and (7) coincide (since by definition $\widetilde{\varphi}(t) = \varphi(t)$ for $t < \alpha$) and, in addition, the first symbols in (3) and (8) coincide, then for $t \leq \alpha$ $\widetilde{\varkappa}(t) = \varkappa(t)$ and, in particular, taking into account $\widetilde{\varkappa}(\alpha) = \varkappa(\alpha) = \varkappa(\beta)$.

It is now easy to see that for arbitrary s the sequence

$$\widetilde{\varkappa}(\alpha) \widetilde{\varkappa}(\alpha+1) \ldots \widetilde{\varkappa}(\alpha+s)$$

coincides with the sequence

$$\varkappa(\beta) \varkappa(\beta+1) \ldots \varkappa(\beta+s),$$

since both of these sequences are the result of the operation of automaton A for identical input sequences

$$p_{\widetilde{\varphi}(\alpha)} p_{\widetilde{\varphi}(\alpha+1)} \cdots p_{\widetilde{\varphi}(\alpha+s)},$$

$$p_{\varphi(\beta)} p_{\varphi(\beta+1)} \cdots p_{\varphi(\beta+s)}$$

and identical initial states $\widetilde{\varkappa}(\alpha) = \varkappa(\beta)$.

Since automaton A represents the function $\varphi(t)$, then for arbitrary p in the sequence (3) $\varkappa(p) \in M$, while by virtue of the above proof, only those symbols are contained in (8) which are contained in (3), it follows that $\aleph(p) \in M$.

This signifies that the event S^{φ} occurs both in the appearance of the sequence (2) and in the appearance of the sequence (7). But since the set $\{S^{\varphi}\}$ contains only one sequence of each given length p, it follows from this that (2) and (7) fully coincide for arbitrary p, i.e.,

$$\widetilde{\varphi}(p) \equiv \varphi(p).$$

For $p < \alpha$, this is automatically satisfied by virtue of (6), while for $p \geq \alpha$ this is necessary, in order that

$$\varphi(p) = \varphi(p + \beta - \alpha).$$

This signifies that the function $\varphi(t)$ is periodic in the last analysis with period $T = \beta - \alpha$, Q.E.D. This completes proof of the theorem.

LITERATURE CITED

1. M. A. Aizerman, L. A. Gusev, L. I. Rozonoér, I. M. Simirnova, and A. A. Tal', Finite Automata. I. Avtomatika i Telemekhanika Vol. 21, No. 2, 1960.
2. M. A. Aizerman, L. A. Gusev, L. I. Rozenoér, I. M. Smirnova, and A. A. Tal', Finite Automata. II. Avtomatika i Telemekhanika Vol. 21, No. 3, 1960.
3. S. C. Kleene, Introduction to Metamathematics [Russian translation] Izd. Inostr. Liter., 1957.
4. R. Péter, Recursive Functions [Russian translation] Izd. Inostr. Liter., 1954.
5. S. C. Kleene, Representation of Events in Nerve Nets and Finite Automata. Automata Studies [Russian translation]. Izd. Inostr. Liter., 1956.
6. I. M. Copi, C. Elcot, and J. V. Wright, Realisation of Events by Logical Nets. Journal Assoc. Comp. Mach. Vol. 5, No. 2, 1958.
7. B. A. Trakhtenbrot, The Synthesis of Logical Networks with Operators Described by Means of the Unary Predicate Calculus [in Russian] Dokl. Akad. Nauk SSSR, Vol. 68, No. 4, 1958.
8. M. Davis, Computability and Unsolvability, McGraw-Hill, New York, 1957.

CANONICAL METHOD OF SWITCHING CIRCUIT SYNTHESIS*

A. Sh. Blokh

(Minsk)

Translated from Avtomatika i Telemekhanika, Vol. 22, No. 6, pp. 756-764, June, 1961
Original article submitted November 17, 1960

A new method for the synthesis of contact networks is described. Upper limits are given for the total and mean numbers of contacts.

The automation of relay control-circuit synthesis requires the development of simple and improved synthesis algorithms.

In this article we present a method for the synthesis of switching circuits which we term canonical. The canonical method for electronic circuits has been described in [1]. The basic concept of the canonical method is the canonical table of Boolean functions.

* The basic results of the article were presented 19 November 1959 at the Section of Cybernetics and Mathematical Logic of the All-Union Conference on Computation and Computers.

We remark that the canonical table permits the unified treatment of the methods of universal multipoles, the method of stages [2], the graphical method of V. N. Roginskii [3], and the construction of twopoles by the method of O. B. Lupanov [4].

Assignment of the conductance function of a two-terminal contact network. The structural conductance of a two-terminal contact network is defined by a Boolean function of binary arguments $y = f(x_1, x_2, \ldots, x_n)$ We shall adopt the following convention for describing Boolean function. We put the set $x_1 = \alpha_1, x_2 = \alpha_2, \ldots, x_n = \alpha_n$ into correspondence with a number $N = \alpha_n + 2\alpha_{n-1} + \ldots + 2^{n-1}\alpha_1$. It is obvious that $0 \leq N \leq 2^n - 1$. Let us define a row of 2^n points. To each point we assign an ordinal number between 0 and $2^n - 1$. Over the point with index $N = \alpha_n + 2\alpha_{n-1} + \ldots + 2^{n-1}\alpha_1$ we write the value of the function $y = f(x_1, x_2, \ldots, x_n)$ for $x_i = \alpha_i$, i.e., $f(\alpha_1, \alpha_2, \ldots, \alpha_n)$. For example, for the function $y = \bar{x}_1 x_2 + x_1 \bar{x}_2$ we obtain the row 0110. We omit the indices of the points, as they are defined by the positions of the points in the row.

Synthesis by the canonical method is carried out in two steps. We first construct the canonical table from the Boolean function, and then realize from it the circuit.

Canonical table. The canonical table for a Boolean function $y = f(x_1, x_2, \ldots, x_n)$ has a triangular shape and contains $(n + 1)$ rows.

The top nth line contains 2^n indices, the bottom, a single index. The top row constitutes the row of the function to be realized. We now describe the method for constructing the canonical table. Assume that the kth row read from below, has already been constructed. It contains 2^k indices. We divide the kth row from left to right into pairs. Below each pair (a, b) in the kth row we put a number c in the $(k-1)$th row, where to different pairs (a, b) we assign different indices c, to identical pairs, identical indices c.

To the pair (a, a) we assign the index a. If in the $(k-1)$th row it is found that $c_2 = c_1$ and c_1 is to the left of c_2, then c_2 is marked in some manner to indicate that in further synthesis c_2 will be ignored.

A table constructed in this manner will be termed canonical. For the function $y = x_1 x_2 + \bar{x}_1 \bar{x}_2$ we have the following table:

1 0 0 1
2 3
4

The two-terminal contact network, corresponding to a canonical table of $(n + 1)$ rows, is realized in the form of a series connection of n blocks r_k, where each block r_k contains the contacts of only a single relay X_k. The indices of the kth row, as the outputs of r_k. To obtain the circuit of r_k it is necessary to connect the index c, defining the input and standing under the pair of indices (a, b), by the contact x_k, to the index (node) b and by the contact x_k to the index (node) a. If a = b (consequently, c = a), the index c is short-circuited to the index a. If a = 0 (b = 0), index c is not connected with a (b). To obtain the outputs of r_k, it is now necessary to connect identical indices (nodes) of the kth row to a single output. The basis for the method is given in the Appendix.

Example 1. Realize the Boolean function

$$y = x_1 x_2 x_3 + x_1 \bar{x}_2 \bar{x}_3 + \bar{x}_1 x_2 \bar{x}_3 + \bar{x}_1 \bar{x}_2 x_3.$$

For this function we obtain the following canonical table:

0 1 1 0 1 0 0 1
2 3 3 2
4 5
6

The top row gives the function itself. The circuit is shown in Fig. 1.

Undefined states. In the canonical method of synthesis it is easy to calculate the number of contacts necessary, from the canonical table. For this it is necessary to determine all distinct indices contained in the table, beginning with the (n−1)th row. Two contacts correspond to the table entries a, b, c, for which $a \neq b$, $ab \neq 0$; a single contact, if $a \neq b$, $ab = 0$; in the remaining cases there are no contacts.

Therefore, if in the table of the function there are undefined states, it is necessary to complete the function $f(x_1, x_2, \ldots, x_n)$ in such a way as to obtain the smallest possible number of different indices in the canonical table. Naturally, we try to have identical indices in the lower rows of the canonical table, since in this case the upper rows of the table will have the smallest number of distinct indices.

Let us give one possible method for completing the function. We divide the top row of the table into two parts, and attempt to complete the function to make these parts identical. We then divide the upper row into four parts and attempt to complete the remaining undefined state to have the smallest number of distinct quarters, then into eight parts, etc.

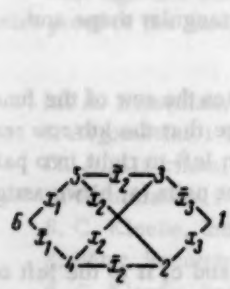


Fig. 1.

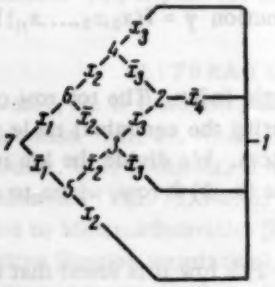


Fig. 2.

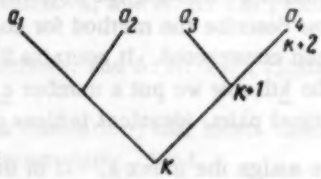


Fig. 3.

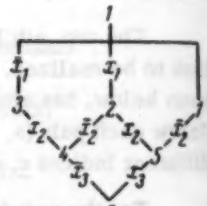


Fig. 4.

Example 2. A Boolean function in the arguments x_1, x_2, x_3, x_4 is given by the row 110 · 1 · 111 · 0 · 011. Dots denote undefined states (constituents). Using the above completion method, we construct the canonical table:

·	1	1	0	·	1	·	1	1	1	·	0	·	0	1	1
1		2		1		1		1		2		2		1	
	3				1				3				4		
			5								6				
							7								

The corresponding circuit is shown in Fig. 2. (In this case index 5 could be short-circuited to index 3 due to the presence of 1 in the pair).

Permutation of arguments. The realization of the twopoles depends on the order selected for $x_1, x_2, \ldots, x_n$, i.e., on the choice of auxiliary function $N = x_n + 2x_{n-1} + \ldots + 2^{n-1}x_1$.

In this function the order of arguments defines the coefficients with which the x_i enter into the function N.

Let us consider the changes in the top row with transposition of the arguments x_{k+1}, x_{k+2}.

Consider one of the nodes in the kth row and the nodes above it in the (k + 1)th and (k + 2)th rows. In the (k + 2)th row we have four nodes (Fig. 3). With transposition of x_{k+1}, x_{k+2} the nodes interchange position, a_2, a_3, since these nodes correspond to the values 1,0 and 0,1 in the arguments of x_{k+1}, x_{k+2}. From this it follows that if the indices of the (k + 2)th row are divided into quarters, corresponding to the indices of the kth row, with transposition of x_{k+1}, x_{k+2} the inner pairs of the quarters exchange positions. Consequently, to obtain the transpositions of the upper, initial row, with transposition of x_{k+1}, x_{k+2}, it is necessary to divide the top row into 2^k large parts (number of indices in the kth row), then divide each part into four smaller parts, and interchange the second and third small parts in each of the larger parts.

In the synthesis method described here, the order of arguments has been strictly observed, which would correspond to circuits consisting of the series connection of individual blocks of contacts from single relays in a defined order. In a certain sense, this could be considered a defect of the method, since it would limit the class of circuits by which a Boolean function is realized. It is true that well-known fairly effective methods for the synthesis of bridge networks [2, 3] have this defect. In this canonical method, it is possible to remove this restriction. For this, it is sufficient to transpose in the top row not all parts, but only individual parts corresponding to individual nodes of the kth row. The validity of this statement follows from the above considerations, illustrated in Fig. 3.

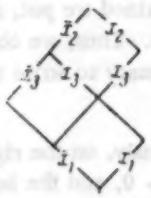


Fig. 5.

This generalization exceeds the limits of the previously considered circuits and can lead to circuits with a smaller number of contacts. In this sense the method is better. But, like each improvement, it complicates the method of partition, which, however, can also be programmed.

Synthesis of (p, 1)-poles. The contacts of a (p, 1)-poles is given by p Boolean functions $y_1 = y_1(x_1, x_2, \ldots, x_n), \ldots, y_p = f_p(x_1, x_2, \ldots, x_n)$. It is necessary to construct such a multipole, that the conductance between the ith input and output be $f_i(x_1, x_2, \ldots, x_n)$. The method of constructing the (p, 1)-pole is analogous to the method of constructing a twopole.

The canonical table is constructed analogously to the above-described case for p = 1. In the top row we write the $p2^n$ values of the p Boolean functions in n arguments. In the bottom, zero row, we obtain p indices, corresponding to the outputs of the circuit. Realization is carried out in the usual manner.

In a particular case, where the f_i are pairwise disjoint, i.e., $f_i f_j = 0$ for $i \neq j$, it is possible to employ a single table. In this case, we put in the top row not 0,1, but the indices of the functions 1, 2, 3, . . . , p and 0. For example, for the functions $y_1 = \bar{x}_1x_2x_3 + x_1\bar{x}_2x_3$, $y_2 = x_1\bar{x}_2\bar{x}_3 + \bar{x}_1x_2\bar{x}_3$, $y_3 = \bar{x}_1\bar{x}_2\bar{x}_3 + x_1x_2x_3 + x_1x_2\bar{x}_3$, we obtain the following table:

3	0	2	1	2	1	3	3
4		5		5		3	
	6				7		
			8				

The table is synthesized in the usual manner.

Introduction of auxiliary arguments. Introduction of an auxiliary argument reduces the volume of the canonical table. This reduces the required memory capacity, which in turn leads to reduction of machine time, and permits the synthesis of circuits with a larger number of arguments. The auxiliary argument can sometimes be found directly from the verbal description of the Boolean function. This, in a certain sense, characterizes the invariant properties of the given Boolean function. For example, for a symmetrical Boolean function it is possible to take as the auxiliary argument the ordinary sum of arguments.

In the general case the finding of an auxiliary argument for a Boolean function is a problem of linear programming.

Let there be given a Boolean function of the auxiliary integral argument $y = f(s)$, $s = \varphi(x_1, x_2, \ldots, x_n)$.

The general principles of synthesizing the circuit do not differ from the above synthesis method: we construct the canonical table and then synthesize it.

Let us give the method of finding the canonical table.

Since s is a function of binary variables, it is representable in the form of a sum of monomials with integral coefficients. Let $s = \sum_{i=1}^{m} a_i\sigma_i$, where the a_i are integral coefficients while the σ_i are monomials in x_i, $\bar{x}_i$. We put $s_k = \sum_{i=1}^{k} a_i\sigma_i$, $s_0 = 0$ $(1 \leqslant k \leqslant m)$. We write in the top row of the table, in increasing order, all values

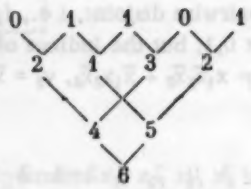
$s_m = s$. Below this row we write the values of s_{m-1}, also in increasing order. In the same way we construct all remaining rows. In the last, lowest row there will be a single value $s_0 = 0$. In the table thus constructed we join each index γ of the kth row by two line segments to the indices γ and $\gamma + a_{k+1}$ in the $(k+1)$th row. Above the indices s_m of the top row we write the corresponding values of the functions. The previous indices s_i can now be struck out; they were necessary only for plotting the line segments. At the nodes of the table obtained we put, according to the above method for constructing the canonical table in the general case, new indices. Thus, we obtain the canonical table. The technique for constructing it is very simple. In fact, it is not necessary to write the indices s_k. It is only necessary to write the values of the coefficients a_k.

The synthesis of the canonical table is basically carried out as before, except that if previously, on the right-hand segments, we marked x_k, and on the left $\bar{x}_k$, we now, on the segments to the right with $a_k > 0$, and the left, with $a_k < 0$, form a series network corresponding to the monomial σ_k, while on segments at the left with $a_k > 0$, and the right with $a_k < 0$, a parallel network, corresponding to $\bar{\sigma}_k$. This method is easily extended to circuits of other logical elements.

Example 3. For the function

$$y = \begin{cases} 1 \text{ for } s = 0, 1, 3, \\ 0 \text{ for } s = -1, 2, \end{cases}$$

where $s = -x_1 + 2x_2 + x_3$, we have the table:



The corresponding circuit is shown in Fig. 4.

The canonical synthesis method, inclusive of partitioning, has been programmed for the "URAL." We have obtained canonical tables for Boolean functions of seven and eight arguments.

On the syntheses of twopoles. We indicate a slight extension of the canonical method for the synthesis of twopoles.

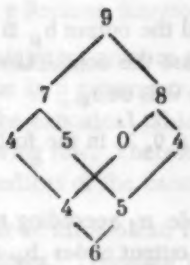
If the canonical table of the Boolean function is divided into two parts along the (n−m)th row, in which there are p different indices, we then obtain in the lower part the canonical table for a (1, p)-pole in (n−m) arguments while in the upper, p "small" tables for the (p, 1)-poles in m arguments. If the "small" tables are pairwise disjoint, i.e., the functions defined by them are pairwise disjoint, then these p tables can be reduced to a single table [see the special case of synthesis of the (p, 1)-pole].

Example 4. Let there be given the canonical Table 1.

1	0	0	1	0	1	0	0	
	2		3		3	0		4 5 0 4
		4			5			7 8
				6				9

The first row defines two disjoint canonical functions with apices 4, 5. We reduce them to the single canonical Table 2. For this we transpose the index of the apex of a given table to the locations of the unities in the top row of the given table. After this the table is constructed in the usual manner.

For the initial function 10010100, we obtain as a result Table 3. The corresponding circuit is represented in Fig. 5.



If only a certain set of "small" tables is pairwise disjoint they can be reduced to a single one synthesized separately from the remaining tables. This extension of the method is efficiently combined with expansion of the initial table.

Expansion of the table is its representation by r such tables, such that the Boolean sum of their top rows be equal to the top row of the given table. This corresponds to representation of a Boolean function by the sum of r Boolean functions.

We now describe a special form of expansion which leads to the construction of twopoles proposed by O. B. Lupanov [4].

The top rows of 2^m indices 0.1 in the "small" canonical tables above the (n−m)th row of the given table are separated into r parts with t indices each, except a possible last part with a smaller number of indices. It is obvious that

$$r = \left[\frac{2^m}{t}\right] + 1.*$$

We expand a given table in r tables $M_1, M_2, \ldots, M_r$. To the top row of M_j we transpose the jth parts of the "small" tables, the remaining indices of the row M_j will be 0. It is obvious that the "small" tables with differing M_j are pairwise disjoint. Therefore "small" tables, each taken from a different M_j, may be reduced to a single table. There will be not more than 2^t such reduced tables, i.e., not more than the maximum number of distinct "small" tables in the individual M_j. Synthesizing the reduced tables of m arguments and r tables in n−m arguments, we obtain a realization of the prescribed Boolean function in the form of the construction proposed by O. B. Lupanov.

APPENDIX

Basis of the canonical method. For the conductance function $y = f(x_1, x_2, \ldots, x_n)$ we introduce n + 1 intermediate integral arguments, satisfying the following relations:

$$\begin{aligned} r_0 &= \text{const}, \\ r_1 &= \varphi_1(r_0, x_1), \\ r_2 &= \varphi_2(r_1, x_2), \\ r_n &= \varphi_n(r_{n-1}, x_n), \\ y &= r_n, \end{aligned} \qquad (1)$$

where the φ_i are single-valued functions.

The two-terminal contact network corresponding to the function $y = f(x_1, x_2, \ldots, x_n)$ is realized in the form of the series connection of n blocks, where each block r_k contains the contacts of only a single relay X_k. Let us consider the synthesis of a single block.

* Here and below the square brackets denote the integral part.

The function $r_k = \varphi_k(r_{k-1}, x_k)$ defines the kth block. Let the function r_{k-1} take on the values $a_1, a_2, \ldots, a_m$, while the function r_k takes on the values $b_1, b_2, \ldots, b_l$. The values $a_1, a_2, \ldots, a_m$ are interpreted as the input nodes of the kth block, while the values $b_1, b_2, \ldots, b_l$, as the output nodes of the kth block.

The conductance between the input a_i and the output b_ν is taken equal to unity with $x_k = \delta_k$, if the equality $b_\nu = \varphi_k(\delta_k, a_i)$ is satisfied; in the contrary case the conductance between the input a_i and the output b_ν is taken equal to zero. The input node to the first block will be r_0.

Since there are two indices in the top row: 0, 1, in the following block there will be two outputs. In fact, the zero output is not realized.

To each node a_i we assign a binary variable α_i according to the following rule: If $r_{k-1} = a_i$, then $\alpha_i = 1$; if $r_{k-1} \neq a_i$, then $\alpha_i = 0$. Analogously, for the output nodes b_ν we introduce the binary numbers β_ν. Let us find the relations between α, β, x_k. We put $r_k = b_\nu$ and list all values of a_i for which the equalities

$$\varphi_k(a_{i_1}, 1) = \ldots = \varphi_k(a_{i_p}, 1) = b_\nu,$$
$$\varphi_k(a_{j_1}, 0) = \ldots = \varphi_k(a_{j_q}, 0) = b_\nu.$$

are valid.

From this it follows that

$$\beta_\nu = x_k(\alpha_{i_1} + \ldots + \alpha_{i_p}) + \bar{x}_k(\alpha_{j_1} + \ldots + \alpha_{j_q}) \tag{2}$$

To construct the connection circuit in the kth block, it is necessary to connect the node b_ν to the nodes $a_{i_1}, \ldots, a_{i_p}$ by the contact x_k, and to connect the nodes $a_{j_1}, \ldots, a_{j_q}$ to the node b_ν through the contact $\bar{x}_k$.

In this realization β_ν is the conductance between the input to the circuit and the node β_ν.

This is obvious for the first block. Let the statement be valid for the α_i. In this case, there follows from relations (2) the validity of the statement for the β_ν as well.

Since the φ_k are single-valued functions, then, taking r_{k-1}, we find only two indices b_{ν_1}, b_{ν_2}, such that $b_{\nu_1} = \varphi_k(a_i, 1)$, $b_{\nu_2} = \varphi_k(a_i, 0)$. Consequently,

$$\alpha_i = x_k\beta_{\nu_1} + \bar{x}_k\beta_{\nu_2}. \tag{3}$$

From this we find that each node a_i is connected by the contacts x_k and $\bar{x}_k$ to not more than two nodes b_{ν_1} and b_{ν_2}, which is the guarantee that sneak paths not be formed.

It is easily seen that the canonical table defines a system of φ_k functions. In fact, the indices of the $(n-k)$th row are the values of the argument r_k and the correspondence between the pair (a, b) in the kth row and the indices c in the $(k-1)$th row signify that $\varphi_k(c, 1) = b$, $\varphi_k(c, 0) = a$. Obviously, there are φ_ν single-valued functions. We note that r_n takes on only two values: 0, 1, while r_0 takes on the single value, given in the lowest row of the table. By the construction of the table, $y = r_n$.

To synthesize the circuits we do not require passage to the functions φ_ν. Synthesis is carried out directly from the canonical table.

Since the indices of the $(k-1)$th row denote the inputs to the kth block, while the indices of the kth row signify the outputs of the kth block, then, employing formula (3), the following rule can be formulated: The node of the $(k-1)$th row standing below the pair (a, b) in the kth row is connected by the contact x_k to the node b, and by the contact $x_k - c$ to the node a. If $a = b$, then $c = a$ and the node c is short-circuited to the node a. Identical indices of the kth row are joined in a single node. If $a = 0$, then c is not connected to a; the same is valid for $b = 0$.

Generalization 1. Let there be in the top row of a canonical table in n arguments not one nonzero index, but q distinct indices. This signifies the case of the (1, q)-pole, when its Boolean functions $f_i(x_1, x_2, \ldots, x_n)$ do not have common constituents, i.e., $f_i(x_1, x_2, \ldots, x_n)f_j(x_1, x_2, \ldots, x_n) = 0$ for $i \neq j$. In this case the construction of the canonical table and its synthesis do not in any way differ from the case of a twopole. This follows immediately from the fact that, beginning with any arbitrary intermediate row of the twopole table, there exists a canonical table and the corresponding circuit of this type.

Generalization 2. For the p Boolean functions $y_1 = f_1(x_1, x_2, \ldots, x_n), \ldots y_p = f_p(x_1, x_2, \ldots, x_n)$ of the (p, 1)-terminal contact network the system of intermediate arguments (1) is constructed analogously to the system (1) for a single Boolean function. In this case, however, r_0 takes on p distinct values and $r_n = z$, where the function $z = F(\tau, x_1, x_2, \ldots, x_n)$ satisfies the relationship $y_i = F(i, x_1, x_2, \ldots, x_n)$. Graphically this system of functions is constructed in the form of p canonical tables for p Boolean functions.

Upper limit for $L_B(n)$. Let $L_B(n)$ be the minimum number of contacts sufficient for the realization by the canonical method of an arbitrary Boolean function in n arguments. Let us consider the canonical table for the Boolean function $f(x_1, x_2, \ldots, x_n)$. In the synthesis of the canonical table, from a set of identical indices in a given row a single index is joined to two indices of the following row. Therefore the number of contacts is equal approximately to $2R_f(n)$, where $R_f(n)$ is the number of distinct indices of the canonical table, but does not exceed this quantity.

Let us estimate $R(n) = \sup R_f(n)$ over the set of all Boolean functions of n variables. We consider the $(n-m)$th row of the table. In this row there are 2^{n-m} indices. Therefore the number of different indices $R(m, n)$ in the $(n-m)$th row is defined by a 2^m-place binary number in the top row. From this it follows that $R(m, n)$ also does not exceed 2^{2^m}, the number of distinct 2^m-place binary numbers. Thus, $R(m, n) < 2^{n-m}$, $R(m, n) < 2^{2^m}$.

It is evident that the number of distinct indices $R(n)$, for arbitrary whole m_1, satisfies the inequality

$$R(n) = \sum_{m=1}^{n} R(m, n) < \sum_{m=1}^{m_1} 2^{2^m} + \sum_{m_1-1}^{n} 2^{n-m}.$$

In the first sum the terms increase as the square. Therefore, the first sum is less than twice the largest term. Consequently,

$$R(n) < 2 \cdot 2^{2^{m_1}} + 2^{n-m_1}.$$

Taking $m_1 = [\log_2(n - \alpha \log_2 n)]$, where $\alpha > 1$, we obtain

$$R(n) < 2\frac{2^n}{n^\alpha} + \frac{2^n}{\frac{1}{2}(n - \alpha \log_2 n)}.$$

Hence

$$L_6(n) < \frac{2^{n+2}}{n^\alpha} + \frac{2^{n+2}}{n - \alpha \log_2 n}.$$

As $n \to \infty$, we have

$$L_6(n) < \frac{2^{n+2}}{n}(1 + \varepsilon) \quad (\varepsilon \to 0).$$

In the case of a (p, 1)-pole we have the inequality:

$$R(m, n) < 2^{2^m}, \quad R(m, n) < p2^{n-m},$$

from which

$$R(n) < 2 \cdot 2^{2^{m_1}} + p2^{n-m_1}$$

Taking $n = [\log_2(n + \log_2 p - \alpha \log_2(n + \log_2 p))]$, we obtain, as $n \to \infty$

$$L_6(p, n) < \frac{p2^{n+2}}{n + \log_2 p}(1 + \varepsilon) \quad (\varepsilon \to 0).$$

For linear $s = a_1 x_1^{\sigma_1} + \ldots + a_n x_n^{\sigma_n}$, $a_i \geq 0$, it is possible to give a better estimate $R(m, n) \leq pS_{n-m}$, in place of the previous $R(m, n) < p2^{n-m}$. As far as concerns the estimate $R(m, n) < 2^{2^m}$, for slowly increasing coefficients a_i we can substitute the improved estimate $R(m,n) < 2^{Sn-Sn-m}$.

Estimate of the mean number of contacts. Let $H_{n,r}$ denote the class of functions of n variables taking on r values equal to 1, and $(2^n - r)$ values equal to 0. It can be assumed that $u = r/2^n$ is the probability that a given constituent enters into the expansion of the function.

Let us divide the row of 2^n values of the Boolean function into 2^{n-m} parts with 2^m binary values in each part. The probability of two such parts coinciding is equal to $P_m = (u^2 + v^2)^{2^m}$, where $v = 1-u$. Consequently, in the (n−m)th row. We shall carry out the calculation as follows. To each index in the row we assign 1 if all indices in the row to the left are distinct from it; in the contrary case we assign 0. Both of these values will have a certain probability. For the first index the probability of 1 is equal to 1. For the second index the probability of 1 is equal to $Q_m = 1 - P_m$, for the third index Q_m^2, etc. Therefore, the mathematical expectation for the number of distinct indices in the (n−m)th row will be

$$M(n, m) = 1 + Q_m + \ldots + Q_m^{2^{n-m}} < \frac{1}{1 - Q_m}.$$

Thus, in the (n−m)th row the mean number of distinct indices is bounded by the number $\frac{1}{(u^2 + v^2)^{2^m}}$. With increasing m this number increases.

But N(n, m) does not exceed the total number of indices in the (n−m)th row. We thus obtain for M(n, m) the two inequalities:

$$M(n, m) < \frac{1}{(u^2 + v^2)^{2^m}}, \qquad M(n, m) < 2^{n-m}.$$

For the mean number of distinct indices M(n) we have, for arbitrary whole m_1, the inequality

$$M(n) = \sum_{m=1}^{n}{}' M(n, m) < \sum_{m=1}^{m_1} \frac{1}{(u^2 + v^2)^{2^m}} + \sum_{m_1+1}^{n}{}' 2^{n-m}.$$

From this it follows that

$$M(n) < 2 - \frac{1}{(u^2 + v^2)^{2^{m_1}}} + 2^{n-m_1}.$$

Let $\frac{1}{u^2 + v^2} = 2^\beta$. Taking $m_1 = \left[\log_2\left(\frac{n}{\beta} - \frac{\alpha}{\beta}\log_2\frac{n}{\beta}\right)\right]$, where $\alpha > 1$, we obtain

$$M(n) < 2 - \frac{2^n}{\left(\frac{n}{\beta}\right)^\alpha} + \frac{\beta \cdot 2^n}{\frac{1}{2}\left(n - \alpha \log_2 \frac{n}{\beta}\right)}.$$

Multiplying by two, we obtain the estimate for the mean number of contacts

$$C(n) < \frac{2^{n+2}}{\left(\frac{n}{\beta}\right)^\alpha} + \frac{\beta 2^{n+2}}{n - \alpha \log_2 \frac{n}{\beta}}.$$

As $n \to \infty$, we have

$$C(n) < \frac{\beta \cdot 2^{n+2}}{n}(1 + \varepsilon),$$

where $\epsilon \to 0$, if

$$C(n) < -\log_2(u^2 + r^2)\frac{2^{n+2}}{n}(1+\varepsilon).$$

Continuing these consideration, we obtain the estimate for the (p, 1)-pole as $n \to \infty$

$$C(p, n) < -\log_2(u^2 + r^2)\frac{p \cdot 2^{n+2}}{n + \log_2 p}(1+\varepsilon).$$

LITERATURE CITED

1. A. Sh. Blokh, Canonical Method for Electronic Circuit Synthesis [in Russian]. In the collected works of the Institute of Machine Working and Automation, Acad. Sci. USSR, 1960.
2. G. N. Povarov, Mathematica-Logical Investigation of the Synthesis of Contact-Networks with Single Input and k Outputs [in Russian]. In Collection "Logical Studies," Izd. AN SSSR, 1959.
3. V. N. Roginskii, Elements of the Structural Synthesis of Relay Control-Circuit Synthesis [in Russian]. Izd. AN SSSR, 1959.
4. O. B. Lupanov, On a Method of Circuit Synthesis [in Russian]. Radiofizika, Vol. 1, 1958.

RELAY-TYPE STATISTICAL AUTOMATA AND CERTAIN METHOD OF THEIR INVESTIGATION

I. I. Perel'man

(Moscow)

Translated from Avtomatika i Telemekhanika, No. 6,
pp. 765-779, June, 1961
Original article submitted October 14, 1960

Current types of automatic relay systems for the statistical quality control of serial production are considered. Methods are given for finding the limiting distributions of the regulated parameter occurring with the use of these systems under conditions of linear time variation of a deterministic perturbation. The possibility of optimal choice of parameters in these systems is demonstrated.

The quality of serially produced objects usually depends on a large number of physicochemical and technological factors having differing degrees of variability in the process of production. The variation of these factors causes deviation of the product quality from the prescribed requirements and special measures are required to eliminate these variations, termed further corrective adjustments of the equipment.

In many serial production processes gauging of the results of a given technical operation on a given sample (part) can be realized only after the operation on the given part is completed. Under these conditions the corrective adjustments to the equipment to obtain the necessary quality of the given part in a given series can be realized on the basis of the results of gauging previously finished parts of the same series.* The possibility of realizing such

*By series we understand the ensemble of identical parts fabricated one after other on the same installation in the absence of explicit changes in the production conditions (for example, without change in the quality of raw materials, without change of gauging instruments, etc.).

a corrective adjustment is based on the fact that in many production processes, together with random quality fluctuations, characteristic only of a given sample, there exist quality variations constituting a regular (deterministic) function of n, where n is the ordinal number of the part in the given series. The presence of such regular components is caused, for example, by progressive wear of the equipment, slow temperature variations, etc.

The magnitude of the regular component cannot usually be given in advance, but its value for a given nth detail can be estimated on the basis of a statistical treatment of the measurements on the preceding parts, belonging to the same series. On the basis of such an estimate it is possible to realize corrective adjustment of the equipment for the purpose of reducing the quality deviations of the nth part from the prescribed nominal value. Following [1], such a procedure is termed statistical quality control, widely used in mass production (cf., for example, [2-4]). However, in these applications, statistical quality control, constituting essentially a nonautomated process, has its principle purpose not in the maximum stabilization of production quality but in reduction of the work involved in sampling quality control. Rising requirements on manufacturing precision have made it necessary to establish automatic statistical quality control systems. The first examples of such systems appearing in the recent years have been employed principally for the corrective adjustment of metal-working machines [5-11], as well as for other production processes [12, 13].

The overwhelming majority of currently known automatic statistical quality control systems are in principle of a single type, which we shall term relay-type statistical automata (RTSA).

1. Relay-Type Statistical Automata

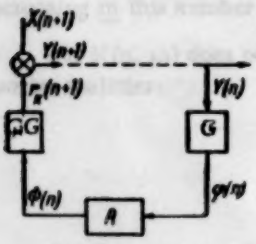


Fig. 1.

To concretize the further discussion, let us consider RTSA employed for the corrective adjustment of a certain technological aggregate, with the aim of stabilizing the geometrical dimensions of the parts produced on it.

Let the controlled dimension of the given (nth) part in a given series have a derivative Y(n) from the prescribed nominal value, which may be represented in the form

$$Y(n) = X(n) + r_k(n) = r(n) + \delta(n) + r_k(n), \tag{1}$$

where X(n) is the deviation of the controlled parameter in the absence of corrective adjustments, r(n) is the regular (deterministic) component of X(n), $\delta(n)$ is the random component of X(n), $r_k(n)$ is the correction introduced by the adjustive mechanism.

Since the quantity Y(n) can be measured only after processing of the nth part is completed, the process of stabilizing the controlled dimension about the value Y = 0 consists in an estimate of the expected value of Y(n) on the basis of the measured values Y : y(n−1), y(n−2), . . ., and introduction of the corrective adjustment $r_k(n)$, in the direction of decreasing the expected value of Y(n). In RTSA this process is carried out in the following way.

In a certain neighborhood of the point Y = 0, corresponding to the nominal value of the regulated parameter (dimension), a zone of admissible positions of the center of scattering (i.e., of permissible values of the mathematical expectation Y) is defined. The position of the center of scattering with respect to its permissible zone is estimated by a special instrument—an analyzer (Fig. 1). The results of measuring already finished parts are applied to the input of the analyzer A from the gauge G. The analyzer distinguishes three states of the center of scattering: positions within the zone, and excursions beyond its positive and negative limits. In this regard, after the part with number n is gauged, a signal $\Phi(n)$ appears at the analyzer output which can have one of three values: 0, +1, -1. The signal $\Phi(n) = 0$ corresponds to the absence of the need for corrective adjustment [$r_k(n+1) = r_k(n)$]. The signals +1 and -1 cause single corrective adjustment steps $r_k(n+1) = r_k(n) + a$ or $r_k(n+1) = r_k(n) - a$, respectively, where the absolute value of the corrective step a is a given constant. Thus, the corrective effect of the adjustment mechanism C is equal to

$$r_k(n) = a\sum_{j=1}^{n-1} \Phi(j). \tag{2}$$

The signal $\Phi(n)$ is a function of a certain set of magnitudes $\varphi(j)$ ($j = n, n-1, n-2, \ldots$), where $\varphi(j)$ is a function of the comparison of the dimension of the jth detail with the control limits, established at preassigned distances $\pm b$ from the prescribed nominal value:

$$\varphi(j) = \begin{cases} 0 & \text{for} \quad |y(j)| \leqslant b, \\ 1 & \text{for} \quad y(j) < -b, \\ -1 & \text{for} \quad y(j) > b. \end{cases} \tag{3}$$

Depending on the method of determining $\Phi(n)$, i.e., on the method of estimating the position of the center of scattering, existing RTSA can be classified among three types.

I. In the simplest case, where the range of production tolerances greatly exceeds the range of scattering of the random deviations, the quantity b may be taken sufficiently large so that, with coincidence of the center of scattering and the prescribed nominal value, the probability of the dimensions of individual parts exceeding the negative (I) or positive (II) control limit of the zone be practically equal to zero (Fig. 2). From this it follows that the first measurement occurring beyond the control limits of the zone should be accompanied by a corrective adjustment, i.e., the equality

$$\Phi(n) = \varphi(n) \tag{4}$$

should be satisfied.

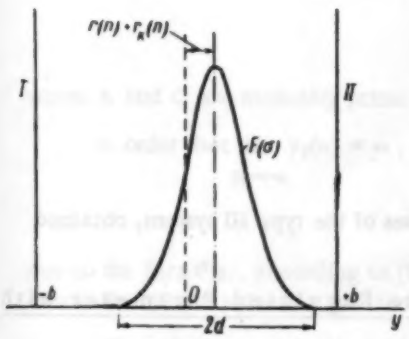


Fig. 2.

The magnitude of the corrective step a in this case can be taken fairly large $b \geq a \geq b-d$, where d is the half-width of the random variation scattering field (cf. Fig. 2).

The above conditions are characteristic, for example, for systems used for the control of centerless grinders [8, 9] and in the experimental machine ÉNIMS [5].

For higher requirements on the precision of the adjustment, i.e., with decrease in the width of the tolerance field to a value close to the width of the random scattering field 2d, the control limits are established within the random scattering field $b < d$ (Fig. 2). Here, obviously, even with coincidence of the center of scattering with the nominal value, i.e., with total compensation of the regular component, there will exist a certain probability of the deviation of the dimensions of individual parts beyond the limits of the control limits, i.e., a probability of incorrect adjustment, leading to increase of the scatter in the stabilized dimension.

It is possible to decrease the influence of incorrect adjustments either by a suitable, choice of magnitude of the adjustment step, respecting relationship (4) (system with "small-pulse" adjustment [5]), or by decreasing the probability of incorrect adjustments by using systems of types II and III.

II. In type II RTSA the corrective adjustment is carried out only when a certain number m of parts in succession exceed the same control limit, i.e.,

$$\Phi(n) = \begin{cases} 1 & \text{for} \quad \sum_{j=n-m+1}^{n} \varphi(j) = m, \\ -1 & \text{for} \quad \sum_{j=n-m+1}^{n} \varphi(j) = -m, \\ 0 & \text{for} \quad \left| \sum_{j=n-m+1}^{n} \right| < m. \end{cases} \tag{5}$$

Here, if $|\Phi(n)| = 1$, after the corrective adjustment is completed, the nth part is not measured, i.e., it is assumed that $\varphi(n) = 0$.

Examples of such systems are given in [7] and [11], where m = 2.

III. In type III RTSA a regular corrective adjustment is carried out in those cases where, among the M last parts produced since the previous corrective adjustment, the difference between the numbers of times that the negative and positive control limits are exceeded reaches or exceeds m, where M and m are preassigned numbers. In other words, if the last corrective adjustment was carried out after measurement of the ith part, with n > i we have:

for $n \geq i + M$

$$\Phi(n) = \begin{cases} \operatorname{sign} G(n), & \text{if} \quad |G(n)| \geqslant m, \\ 0 & \text{if} \quad |G(n)| < m, \end{cases} \tag{6}$$

for $n < i + M$

$$\Phi(n) \equiv 0,$$

where

$$G(n) = \sum_{j=n-M+1}^{n} \varphi(j).$$

An example of this type of RTSA is given in [12].

We remark that type I and II systems may be considered as particular cases of the type III system, obtained with M = n and M = m = 1, respectively.

2. Investigation of RTSA Operation in Stabilization of the Regulated Parameter with Linearly Increasing Perturbation

Let us consider operation of RTSA within the limits of a series consisting of a certain number of parts n_k. We shall terms the instant of termination of the nth part the nth instant of time ($n = 1, 2, \ldots, n_k$), and we adopt the following assumption.

A. The regular component X(n) varies linearly, i.e.,

$$X(n) = cn + \delta(n), \tag{7}$$

where c is constant for the given series.

Equation (7) may be considered characteristic for the operation of a series of metal-cutting machines, where the term cn is due to the progressive wear of the cutting instrument.

B. The distribution of random deviations is independent of n, and there is no correlation between the random deviations of two different parts.

Obviously, the mathematical expectation of the random deviations is equal to zero, since in the contrary case it would have to be taken into account in the regular component.

C. The integral distribution of random deviations is given by a piecewise-continuous function F(δ). Here there exists a magnitude 2d (width of scattering range of the random deviations), such that

$$F(-d) = 0, \quad F(d) = 1, \quad 0 < F(\delta) < 1 \quad \text{for} \quad |\delta| < d. \tag{8}$$

D. The quantities c and d characterizing the perturbation X(n), and the quantities a and b, characterizing the parameters of the automaton, are given whole numbers. For rational a, b, c and d this can always be obtained by a suitable choice of scale for the regulated parameter Y.

We shall demonstrate that with these assumptions the distribution of the set of quantities Y(n) (n = 1, 2,...,n_k) tends as $n_k \to \infty$ to the function $\Psi(y)$, which may be calculated on the basis of prescribed a, b, c and $F(\delta)$. We shall carry out the proof for the type III RTSA with prescribed values of M and m. At the same type all relations obtained will be valid for type I and II RTSA with the substitutions M = m and M = m = 1, respectively.

Considering (1), (2) and (7), we find the mathematical expectation of the quantity Y(n):

$$y_0(n) = M[Y(n)] = cn + a\sum_{j=1}^{n-1}\Phi(j) = cn + an', \tag{9}$$

where n' is equal to the difference between the numbers of positive and negative corrective adjustments carried out up to the time n($|n'| < n$).

The quantity $y_0(n)$ will be termed the position of the scattering center at the nth instant of time.

From (9) it follows that all possible values of $y_0(n)$ must be whole numbers, multiples of α, where α is the largest common divisor of a and c, such that

$$a = \alpha A \quad \text{and} \quad c = \alpha C, \tag{10}$$

where A and C are mutually prime whole numbers.

In order that $\lim_{n\to\infty} y_0(n) \neq \infty$, the magnitude of the corrective step should satisfy the relationship

$$a \geqslant Mc, \tag{11}$$

due to the fact that, according to (6) and (9), during the interval between two corrective adjustments the scattering center shifts by a value not greater than Mc.

Where (11) is satisfied, which shall be assumed everywhere below, the possible values of $y_0(n)$ are bounded: $\bar{Y}_{min} \leq y_0(n) \leq \bar{Y}_{max}$ (n = 1, 2, . . . , ∞), where $\bar{Y}_{min}$ and $\bar{Y}_{max}$ are finite quantities whose values can be found from the given process parameters (cf. Appendix I). Thus, the number of all possible positions of the scattering center in the given system is equal to

$$N = \frac{\bar{Y}_{max} - \bar{Y}_{min}}{\alpha} + 1. \tag{12}$$

The coordinates of the N possible positions along the Y axis are defined by the quantity

$$\bar{Y}(s) = \bar{Y}_{min} + (s-1)\alpha, \tag{13}$$

where s is the ordinal number of the position (s = 1, 2, . . . , N).

If it be given that $y_0(n) = \bar{Y}(s)$, then, on the basis of (8) and (9), the probability of all possible values of $\varphi(n)$ and $y_0(n+1)$ are defined by the system of equations (14) and (15), respectively:

$$\begin{aligned}
&P[\varphi(n) = +1] = P[\delta < -b - \bar{Y}(s)] = F[-b - \bar{Y}(s)],\\
&P[\varphi(n) = 0] = P[-b - \bar{Y}(s) \leqslant \delta \leqslant b - \bar{Y}(s)] =\\
&\qquad = F[b - \bar{Y}(s) + 0] - F[-b - \bar{Y}(s)],\\
&P[\varphi(n) = -1] = P[\delta > b - \bar{Y}(s)] = 1 - F[b - \bar{Y}(s) + 0]
\end{aligned} \tag{14}$$

and

$$\begin{aligned}
&P[y_0(n+1) = \bar{Y}(\delta) + c] = P[\Phi(n) = 0],\\
&P[y_0(n+1) = \bar{Y}(\delta) + c + a] = P[\Phi(n) = 1],\\
&P[y_0(n+1) = \bar{Y}(\delta) + c - a] = P[\Phi(n) = -1].
\end{aligned} \tag{15}$$

The state of the system at the instant $\underline{n}$ will be characterized by a vector with components $y_0(n)$ and $t(n)$, where $t(n)$ is defined by the difference between given instants of time $\underline{n}$ and the instant of carrying out the last corrective adjustment $i (i > n)$:

$$t(n) = n - i \quad \text{for} \quad 1 \leqslant n - i \leqslant M + 1,$$
$$t(n) = M + 1 \quad \text{for} \quad n - i \geqslant M + 1. \tag{16}$$

Thus there exist in all $k = N(M + 1)$ possible states, to each of which we assign its ordinal number $h (h = 1 \text{ to } k)$.

Let the state of the regulated process at the instant $\underline{n}$ be given: $t(n) = T$, $y_0(n) = \overline{Y}(s)$. Then the right-hand sides of Eq. (15) are independent of $\underline{n}$ and can be found as functions of $\overline{Y}(s)$, T and the prescribed process characteristics. Considering (15) and (16), and the fact that corrective adjustments do not occur in the RTSA for $T < M$, we obtain

for $T < M$

$$P[y_0(n+1) = \overline{Y}(s) + c,\ t(n+1) = M + 1] = 1, \tag{17}$$

for $T \geq M$

$$P[y_0(n+1) = \overline{Y}(s) + c,\ t(n+1) = M + 1] = z_1(s, T),$$
$$P[y_0(n+1) = \overline{Y}(s) + c + a,\ t(n+1) = 1] = z_2(s, T), \tag{18}$$
$$P[y_0(n+1) = \overline{Y}(s) + c - a,\ t(n+1) = -1] = z_3(s, T).$$

Here $z_1(s, T) + z_2(s, T) + z_3(s, T) = 1$, and, consequently, no other states except those given by (18) can exist. The method for determining the right-hand sides of the system (18) is given in Appendix II.

On the basis of (17) and (18) a quantity $p_{\xi\eta}$ can be found for arbitrary pairs of possible system states with numbers $h = \xi$ and $h = \eta$, independent of $\underline{n}$. This quantity is equal to the probability that the system existing at a certain instant of time $\underline{n}$ in the state $h = \xi$ pass in the following $(n + 1)$th instant of time to the state $h = \eta$. Thus, the process of variation of the states of the RTSA forms a simple homogeneous Markov chain with a finite number of states $\underline{k}$ and discrete time. The law of formation of this chain [14] is defined by the matrix

$$(p_{\xi\eta}) = \begin{pmatrix} p_{11} p_{12} \cdots p_{1k} \\ p_{21} p_{22} \cdots p_{2k} \\ \cdots\cdots\cdots \\ p_{k1} p_{k2} \cdots p_{kk} \end{pmatrix}. \tag{19}$$

It can be shown that the matrix (19) has no isolated subgroups of states, i.e., that all states existing in the given chain can be found in a single nondecomposable submatrix of matrix (19).

Let observations be carried out on the process satisfying (19) in the time interval from 0 to n_k. Then, as is well known, the ratio of the number of times the process is incident on a certain state $h (h = 1 \text{ to } k)$ to the total number of observation n_k tends as $n_k \to \infty$ to the limit f_h, independently of the initial state of the system. The quantity f_h can be calculated on the basis of the given matrix (19).

Let us denote the characteristic determinant of matrix (19) as

$$P(1) = |E - P|, \tag{20}$$

where E is the unit matrix of order $\underline{k}$, while $\underline{p}$ is matrix (19).

The principal minor of determinant $p(1)$, obtained by suppressing the line and column with index $\underline{i}$, we denote as $P_{ii}(1)$, and the above-mentioned quantity f_h is calculated from the relationship [14]

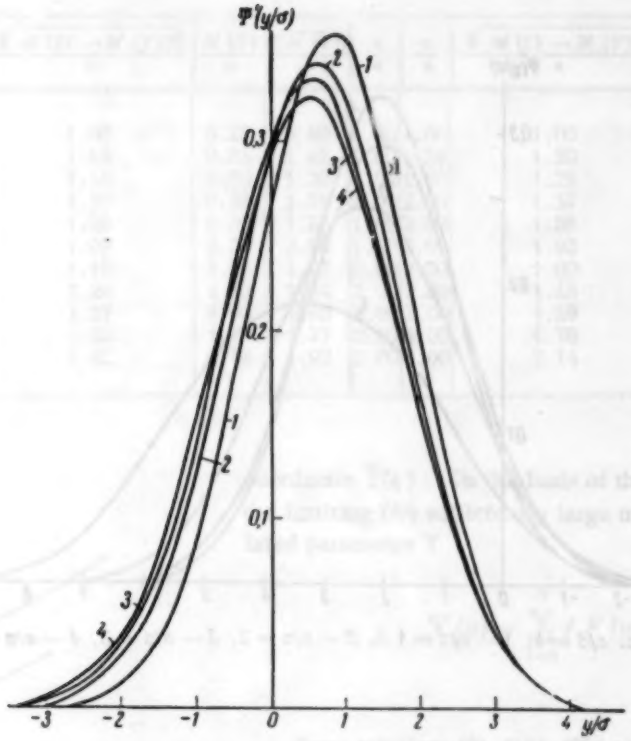


Fig. 3a. $c/\sigma = 0.25$; *1*— $a/\sigma = 0.5$, *2*— $a/\sigma = 0.75$, *3*— $a/\sigma = 1$, *4*— $a/\sigma = 1.25$

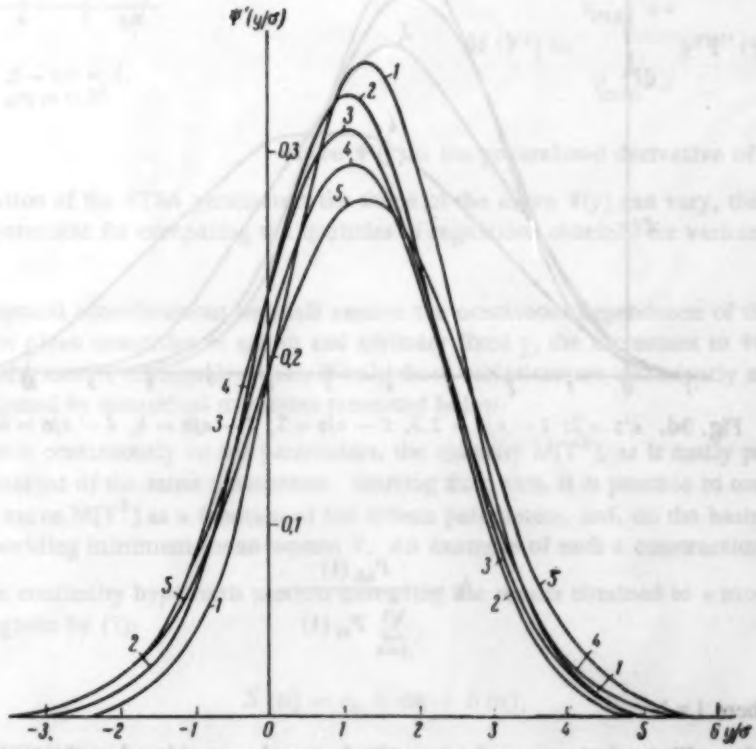


Fig. 3b. $c/\sigma = 0.5$; *1*— $a/\sigma = 0.75$, *2*— $a/\sigma = 1$, *3*— $a/\sigma = 1.5$, *4*— $a/\sigma = 2$, *5*— $a/\sigma = 2.5$

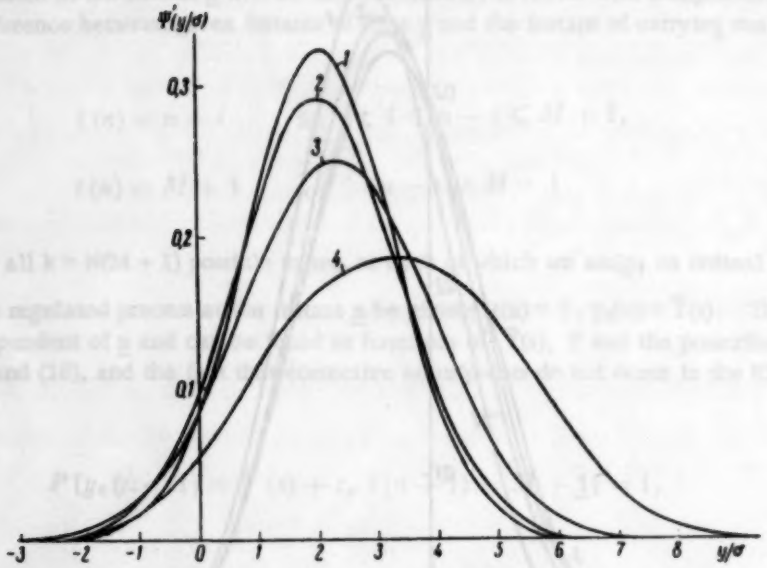


Fig. 3c. $c/\sigma = 1$; *1* — $a/\sigma = 1.5$, *2* — $a/\sigma = 2$, *3* — $a/\sigma = 3$, *4* — $a/\sigma = 5$

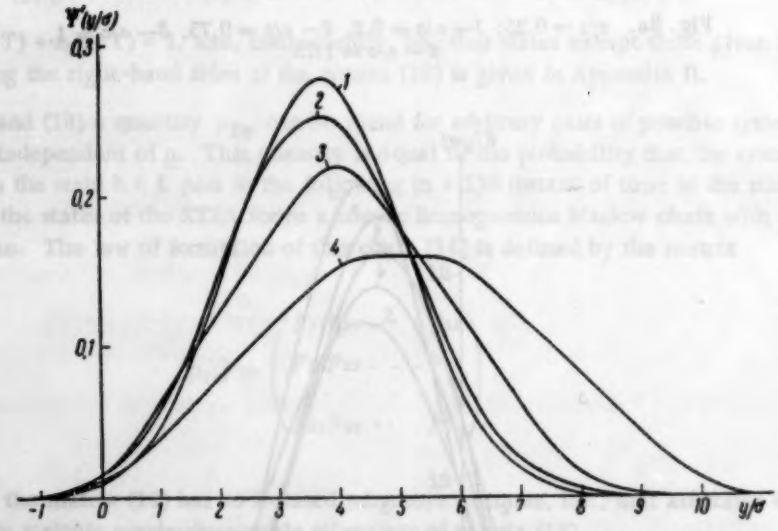


Fig. 3d. $c/\sigma = 2$; *1* — $a/\sigma = 2.5$, *2* — $a/\sigma = 3$, *3* — $a/\sigma = 4$, *4* — $a/\sigma = 6$

$$f_h = \frac{P_{hh}(1)}{\sum_{i=1}^{k} P_{ii}(1)}, \tag{21}$$

where $P_{hh}(1) = P_{ii}(1)$ where $i = h$.

It is obvious that for sufficiently large n_k the quantity f_h may be considered as the probability of finding the system in the state $\underline{h}$. Summing all f_h relating to a state with a given value $\overline{Y}(s)$ (summing over all T from 1 to M + 1), we can find the quantity f_s(s = 1 to N)—the probability of finding the scattering center at the point with

$\frac{c}{\sigma}$	$\frac{a}{\sigma}$	$\frac{\sqrt{M\{[Y-M[Y]]^2\}}}{\sigma}$	$\frac{M[Y]}{\sigma}$	$\frac{\sqrt{M[Y^2]}}{\sigma}$	$\frac{c}{\sigma}$	$\frac{a}{\sigma}$	$\frac{\sqrt{M\{[Y-M[Y]]^2\}}}{\sigma}$	$\frac{M[Y]}{\sigma}$	$\frac{\sqrt{M[Y^2]}}{\sigma}$
0.25	0.25	1.00	3.25	3.40	1.00	1.00	1.00	4.00	4.13
0.25	0.50	1.13	0.85	1.42	1.00	1.25	1.20	2.25	2.56
0.25	0.75	1.16	0.60	1.30	1.00	1.50	1.25	2.05	2.40
0.25	1.00	1.20	0.53	1.31	1.00	2.00	1.37	2.00	2.43
0.25	1.25	1.23	0.50	1.33	1.00	3.00	1.58	2.35	2.84
0.50	0.50	1.00	3.50	3.65	1.00	5.00	1.92	3.30	3.81
0.50	0.75	1.15	1.35	1.77	2.00	2.00	1.00	5.00	5.10
0.50	1.00	1.20	1.12	1.64	2.00	2.50	1.45	3.70	3.98
0.50	1.50	1.27	1.05	1.65	2.00	3.00	1.59	3.65	3.99
0.50	2.00	1.35	1.15	1.77	2.00	4.00	1.76	4.00	4.37
0.50	2.50	1.42	1.35	1.92	2.00	6.00	2.14	5.00	5.45

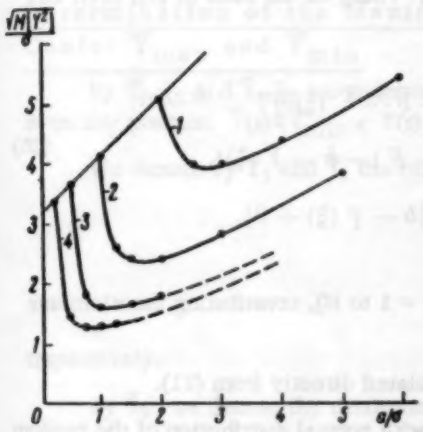


Fig. 4. 4. *1* — $c/\sigma = 2$, *2* — $c/\sigma = 1$, *3* — $c/\sigma = 0.5$, *4* — $c/\sigma = 0.25$

coordinate $\overline{Y}(\delta)$. On the basis of the values of f_s obtained we find the limiting (for sufficiently large n_k) distribution law for the regulated parameter Y

$$\Psi(y) = \sum_{s=1}^{N} f_s F[y - \overline{Y}(s)]. \tag{22}$$

From relations (8), (13), (22) and the piecewise continuity of the function $F(\delta)$, it follows that $\Psi(y)$ is piecewise continuous and $\Psi(\overline{Y}_{min} - d) = 0$, $\Psi(Y_{max} + d) = 1$. From this the mean-square value of Y is defined as

$$M[Y^2] = \int_{\overline{Y}_{min}-d}^{Y_{max}+d} y^2 \Psi'(y)\, dy, \tag{23}$$

where $\Psi'(y)$ is the generalized derivative of $\Psi(y)$ [15].

Since with variation of the RTSA parameters the shape of the curve $\Psi(y)$ can vary, the quantity $M[Y^2]$ is the most convenient characteristic for comparing the qualities of regulation obtained for various values of the RTSA parameters.

From obvious physical considerations we shall assume the continuous dependence of the function $\Psi(y)$ on the system parameters: For given quantities M and m and arbitrary fixed y, the increment to $\Psi(y)$ due to variation in the parameters a, b and c can be arbitrarily small, if only these variations are sufficiently small. The validity of this hypothesis is confirmed by numerical examples presented below.

Since $\Psi(y)$ depends continuously on the parameters, the quantity $M[Y^2]$, as is easily proven on the basis of (23), is also a continuous function of the same parameters. Starting from this, it is possible to construct, by graphical-analytic methods, the curve $M[Y^2]$ as a function of the system parameters, and, on the basis of this curve, find the optimal parameters providing minimum mean-square Y. An example of such a construction is given below.

We note that the continuity hypothesis permits extending the results obtained to a more general case of variation of X(n) than that given by (7):

$$X(n) = c_0 + cn + \delta(n), \tag{24}$$

where c_0 is a given constant.

In effect, let $\Psi(y)$ and $\Psi_1(y)$ be the limiting distributions of Y with $c_0 = 0$ and $c_0 \neq 0$, respectively, and the system parameters a and c correspond to (10). Then, if c_0 is a multiple of α, c(coincides with one of the possible positions of the scattering center obtained for $c_0 = 0$.

Consequently, the behavior of the system for X(n) given by Eq. (7) and Eq. (24) differs only in the initial states which, as we know, have no influence on the limiting distribution and, therefore, $\Psi(y) \equiv \Psi_1(y)$.

Let now c_0 not be a multiple of α. We then consider a new system, differing from the previous one only in that the quantity a is substituted by $a_1 = a + \Delta a$. It is easily shown that the quantity Δa can be taken such that c_0 coincide with one of the possible positions of the scattering center in this new system, while at the same time Δa will be less than an arbitrarily small positive number ϵ. From this, on the basis of the continuity hypothesis, we come to the conclusion that for arbitrary c_0 the limiting distribution of Y coincides with $\Psi(y)$.

3. The Operation of Type I RTSA

Type I RTSA may be studied on the basis of the above general relations, substituting M = m = 1. The general relations can be substantially simplified, since the right-hand sides of (18), as is easily seen, do not depend on T in the present case. Thus, the state of the system at the time n is fully characterized by a single quantity – the position of the scattering center $y_0(n)$. We ascribe the index s to this state, corresponding to the value of s in the relationship $y_0(n) = \overline{Y}(s)$ (s = 1 to N).

Let at the instant of time n the system be in the state with index $s = \xi$. Then, on the basis of (4), (14), and (15) we have

$$\begin{gathered} P\left[y_0(n+1) = \overline{Y}\left(\xi + \frac{c+a}{\alpha}\right)\right] = F\left[-b - \overline{Y}(\xi)\right], \\ P\left[y_0(n+1) = \overline{Y}\left(\xi + \frac{c}{\alpha}\right)\right] = F[b - \overline{Y}(\xi) + 0] - F[-b - \overline{Y}(\xi)], \\ P\left[y_0(n+1) = \overline{Y}\left(\xi + \frac{c-a}{\alpha}\right)\right] = 1 - F[b - \overline{Y}(\xi) + 0]. \end{gathered} \tag{25}$$

On the basis of (25) we calculate the probabilities of passage $p_{\xi\eta}$ (ξ, η = 1 to N), constituting the elements of the matrix (19), having in this case order k = N.

The limiting probabilities f_s in this case coincide with f_h and are calculated directly from (21).

We present below the results of applying this method to a type I RTSA with normal distribution of the random deviations, approximated by the function $F(\delta)$:

$$\begin{gathered} F(\delta) = \frac{1}{\sigma\sqrt{\pi}} \int_{-\infty}^{\delta} e^{-\frac{\delta^2}{2\sigma^2}} d\sigma \qquad \text{for} \quad |\delta| < 3\sigma, \\ F(-3\sigma) = 0, \qquad F(3\sigma) = 1. \end{gathered} \tag{26}$$

The limiting differential distributions of Y calculated* for various values of the parameters a and c, with the condition b = a, are given in Fig. 3a - 3d.

The unit of measurement for all quantities was taken as the root-mean-square deviation of the random component σ. As is evident from Fig. 3a-3d, the curves $\Psi(y)$ for relatively small values of a/σ are very close to the normal distribution. With increase of a/σ the shape of the curves $\Psi'(y)$ gradually changes.

On the basis of the curves of $\Psi'(y)$ found, values of the quantities $\sqrt{M[Y^2]}$, were calculated and tabulated, and used for the plotting of the curve family in Fig. 4. The curves terminate at the left at the locus of minimum admissible values, whose equation

$$\frac{\sqrt{M[Y^2]}}{\sigma} = \frac{\sqrt{(a+3\sigma)^2 + \sigma^2}}{\sigma} \tag{27}$$

is obtained from the fairly obvious circumstance that with a = c the curve $\Psi(y)$ is obtained from the curve $F(\delta)$

*The quantities f were calculated on an electronic computer.

[Eq. (26)] by shifting the latter along the axis of abscissae by the magnitude $a + 3\sigma$. The presence of the line (27) simplifies interpolation of the curve family of Fig. 4 for intermediate values of $\underline{c}$.

The curve family in Fig. 4 can be used both for analysis of the system operating quality with given $\underline{a}$, and for choice of optimum values of $\underline{a}$ under given operating conditions.

Let, for example, the quantities $\underline{c}$ and σ, constant in each given series, vary independently of each other from series to series in the limits: $\sigma_1 \leqslant \sigma \leqslant 2\sigma_1$, $-1{,}5\sigma_1 \leqslant c \leqslant \sigma_1$, where σ_1 is a given constant. It is required to find the value of $\underline{a}$ giving a minimax of the mean-square of Y, i.e., a minimum value of $M[Y^2]$ for the worst series. From an examination of the curves of Fig. 4, it is evident that other conditions being equal, $M[Y^2]$ increases with increase of $|c|$ and σ. Thus, the worst series corresponds to $\sigma = 2\sigma_1$ and $|c| = 1.5\sigma_1$; here $|c|/\sigma = 0.75$. From a consideration of Fig. 4 we conclude that the minimum $M[Y^2]$ with $c/\sigma = 0.75$ corresponds to $a/\sigma = 1.3$. Thus, $a_{opt} = 1.3\sigma = 2.6\sigma$, and the value of $\sqrt{M[Y^2]}$ obtained is equal to $4\sigma_1$.

At the same time, $\sqrt{M[Y^2]} \leqslant 4\sigma_1$ for given $\underline{c}$ and σ, lying in the above limits.

APPENDIX

Determination of the Maximum and Minimum Possible Positions of the Scattering Center $\overline{Y}_{max}$ and $\overline{Y}_{min}$

By $\overline{Y}_{max}$ and $\overline{Y}_{min}$ we understand effective positions of the scattering center, i.e., such as can be reached from any position $\overline{Y}(s)$ ($\overline{Y}_{min} < Y(s) < \overline{Y}_{max}$) in a finite time segment.

We denote by $\overline{Y}_1$ and $\overline{Y}_2$ the maximum values $\overline{Y}(s)$, satisfying the conditions:

$$P[\varphi(n) = 0 / y_0(n) = \overline{Y}(s)] \neq 0, \tag{28}$$

$$P[\varphi(n) = 1 / y_0(n) = \overline{Y}(s)] \neq 0, \tag{29}$$

respectively.

By $\overline{Y}_3$ we denote the minimum value of $\overline{Y}(s)$, satisfying the condition

$$P[\varphi(n) = -1 / y_0(n) = \overline{Y}(s)] \neq 0. \tag{30}$$

Let

$$d + b = s_1\alpha + e_1, \qquad d - b = s_2\alpha + e_2, \tag{31}$$

where s_1, s_2, e_1, e_2 are integers, and $1 \leq e_1, e_2 \leq \alpha$

Now, on the basis of (8) and (14) we obtain

$$\overline{Y}_1 = s_1\alpha = d + b - e_1, \qquad \overline{Y}_2 = s_2\alpha = d - b - e_2, \qquad \overline{Y}_3 = -s_3\alpha = b - d + e_2. \tag{32}$$

The maximum value of $\overline{Y}(s)$ reached from the position $\overline{Y}_1$ up to intervention of a negative corrective adjustment will be denoted as $\overline{Y}_{max\,1}$. On the basis of (6), (9), (16) and (28), we obtain

$$y_0(n) = \overline{Y}_{max\,1} = \overline{Y}_1 + Mc \qquad \text{for} \qquad t(n) = M. \tag{33}$$

The maximum value of $\overline{Y}(s)$ reached by a positive corrective adjustment will be denoted as $\overline{Y}_{max}$. Obviously this will be reached at a certain time instant $n = i + M + 1$, such that

$$y_0(i + m) = \overline{Y}_2, \qquad t(i + m) = m, \qquad \sum_{j=i-1}^{i+m} \psi(j) = m.$$

Thus,

$$y_0(n) = \bar{Y}_{\max 2} = \bar{Y}_2 + (M + 1 - m)\,c + a. \tag{34}$$

Taking into account that $b \geq a$ always, on the basis of (32) - (34), we obtain $\bar{Y}_{\max 1} > \bar{Y}_{\max 2}$, and therefore

$$\bar{Y}_{\max} = \bar{Y}_{\max 1} = d + b + Mc - e_1. \tag{35}$$

The position $\bar{Y}_{\min}$ reached by a negative corrective adjustment at the time instant $n = i + m$ is such that:

$$y_0(i) = \bar{Y}_3, \qquad t(i) = M - m + 1, \qquad \sum_{i=n-M}^{n-1} \varphi(j) = -m.$$

Thus,

$$y_0(n) = \bar{Y}_{\min} = \bar{Y}_3 + mc - a \quad \text{for} \quad t(n) = 1. \tag{36}$$

From this, taking into account (32), we obtain

$$\bar{Y}_{\min} = -a - d + b + mc + e_2. \tag{37}$$

APPENDIX II

Method of Calculating the Probabilities of Passage for the General Case (Type III RTSA)

Let us consider the behavior of the system at the time n such that the last corrective adjustment occurred at the time i($i < n$), while there were no adjustments at the times j($i < j \leq n$).

Let $y_0(n) = \bar{Y}(s)$, then, taking into account (9), we obtain

$$y_0(j) = y_0(n - \tau) = \bar{Y}(s) - \tau c, \tag{38}$$

where $\tau = n - j$ ($i < j \leq n$).

Let us introduce the following notation:

$$p_j = P[\varphi(j) = 1], \qquad q_j = P[\varphi(j) = -1], \qquad r_j = P[\varphi(j) = 0] = 1 - p_j - q_j. \tag{39}$$

On the basis of (15) and (38) we can calculate

$$p_j = p[s, \tau], \qquad q_j = q[s, \tau], \qquad r_j = r[s, \tau]. \tag{39a}$$

We find the probability Q(v, w), such that in the ensemble of M parts with indices $j(n - M + 1 \leq j \leq n)$ there are just v parts with $\varphi = 1$ and just w parts with $\varphi = -1$

$$Q(v, w) = \sum_{\alpha} \prod(\alpha_1, \alpha_2, \ldots, \alpha_v) \sum_{\beta(\alpha)} \prod(\beta_1, \beta_2, \ldots, \beta_w)\, \lambda(\alpha, \beta). \tag{40}$$

Here $\prod(\alpha_1, \alpha_2, \ldots, \alpha_v)$, is the product v of numbers p_j with indices $j = \alpha_1, \alpha_2, \ldots, \alpha_v$, representing one of the $\binom{v}{M}$ possible combinations of all possible values j ($n - M + 1 \leq j \leq n$); $\prod(\beta_1, \beta_2, \ldots, \beta_w)$ for given values $\alpha_1, \alpha_2, \ldots, \alpha_v$, is the product w of numbers q_j with indices $j = \beta_1, \beta_2, \ldots, \beta_w$, representing one of the $\binom{w}{M-v}$ possible combinations of all j, except $j = \alpha_1, \alpha_2, \ldots, \alpha_v$;

$$\lambda(\alpha, \beta) = \lambda(\alpha_1, \alpha_2, \ldots, \alpha_v, \beta_1, \beta_2, \ldots, \beta_w) = r_{\gamma_1} r_{\gamma_2} \ldots r_{\gamma_{(M-v-w)}},$$

where $\gamma_1, \gamma_2, \ldots, \gamma_{(M-v-w)}$ correspond to all possible values of j except $j = \alpha_1, \alpha_2, \ldots, \alpha_v, \beta_1, \beta_2, \ldots, \beta_w$; $\sum_{\alpha}$ signifies summation over all $\binom{v}{M}$ combinations of α; $\sum_{\beta(\alpha)}$ denotes summation over all $\binom{w}{M-v}$ combinations of β, possible for given $\alpha_1, \alpha_2, \ldots, \alpha_v$.

Thus, Q(v, w) is the sum of $\binom{v}{M}\binom{w}{M-v} = \frac{M!}{v!\,w!\,(M-w-v)!}$ terms of the form

$$\prod(\alpha_1, \alpha_2, \ldots, \alpha_v) \prod(\beta_1, \beta_2, \ldots, \beta_w)\,\lambda(\alpha, \beta).$$

Let us find the probability that the quantity G(n), defined according to (6), is equal to an integer L($0 \leq |L| \leq M$):

$$R(L) = P[G(n) = L].$$

It is obvious that the probability R(L) is equal to the sum Q(v, w), taken for all values of w and v satisfying the conditions:

$$v - w = L, \qquad w + v \leqslant M. \tag{41}$$

From this, with $L > 0$

$$R(L) = \sum_{v=L}^{V_{max}} Q(v,\ v-L), \tag{42a}$$

where V_{max} is the largest integer $\leq \frac{M+L}{2}$

With $L < 0$

$$R(L) = \sum_{v=L}^{-W_{max}} Q(w-|L|,\ w), \tag{42b}$$

where $W_{max} = V_{max}$.

On the basis of (39a) and (40) the quantity R(L) can be represented in the form of an explicit function of s, independent of n

$$R(L) = R(L, s).$$

On the basis of (6), for $n > i + M$, we have

$$|G(n) - G(n-1)| = |\varphi(n) - \varphi(n-M)| \leqslant 2.$$

From this it follows that, with $n > i + M$

$$\begin{aligned} |G(n)| &\leqslant m \quad \text{for} \quad G(n)\,\varphi(n-M) \geqslant 0, \\ |G(n)| &\leqslant m+1 \quad \text{for} \quad G(n)\,\varphi(n-M) < 0, \end{aligned} \tag{43a}$$

since in the opposite case the corrective adjustment would have occurred at the instant j, $i + M \leq j < n$.

For $n = i + M$

$$|G(n)| \leqslant M. \tag{43b}$$

Taking into account (6), (42) and (43), we find the values of the probabilities of passage $z_2(s, T)$ and $z_3(s, T)$ in (18) (T = M and T = M + 1).

With T = M (n = i + M)

$$z_2(s, M) = P[G(n) \geqslant m] = \sum_{L=m}^{N} R(L, s), \tag{44a}$$

$$z_3(s, M) = P[G(n) \leqslant -m] = \sum_{L=-m}^{-M} R(L, s). \tag{44b}$$

With T = M + 1 (n > i + M)

$$z_2(s, M+1) = P[G(n) = m] + q_{n-M} P[G(n) = m+1] = \tag{45a}$$
$$= R(m, s) + q(s, M) R(m+1, s),$$

$$z_3(s, M+1) = P[G(n) = -m] + p_{n-M} P[G(n) = -m-1] = \tag{45b}$$
$$= R(-m, s) + p(s, M) R[-m-1, s].$$

LITERATURE CITED

1. W. Shewhart, Economic Control of Quality of Manufactured Products, New York, 1931.
2. Scientific Memoirs on Statistics, Vol. II. Statistical Methods of Industrial Research [in Russian] Izd. AN SSSR, 1954.
3. S. S. Lipkina, V. V. Glovinskii, and G. A. Bryanskii, Experience in the Introduction of Statistical Quality Control Methods in the "Kaliber" Plant [in Russian] Tsentr. Byuro Tekhn. Informatsii, 1952.
4. V. V. Golovinskii, Statistical Quality Control Abroad [in Russian] Mashgiz, 1957.
5. A. I. Al'tshuler, Automatic Dimensional Control Using Small Impulses in Metal Cutting. Automation of Technology in the Metal-Working Industries. Metal Cutting and General Problems of Automation [in Russian] Izd. AN SSSR, 1956.
6. H. G. Amber and P. S. Amber, Developments in Automatic Quality Control. Tool Engineer No. 1, 1956.
7. D. N. Smith, Role of Statistical Computation in Machine Tool Feedback Gauging. Control Eng. No. 9, 1957.
8. J. Loxhem, Control of Quality in Automation. Metalworking Production No. 2, August, 1957.
9. J. Loxhem, The Control of Quality in Automation. Machinery No. 9, August, 1957.
10. A. Parnes, Magnetostriction Actuates Size Controlled Grinders. Automation Progress Vol. 3, No. 10, October, 1958.
11. Simple Counter Detects Machine Drift. An Electrical Manufacturing Staff Report. Electrical Manufacturing Vol. 60, No. 4, October, 1957.
12. P. Huggins, Statistical Computers as Applied to Industrial Control. Journal of British Inst. of Radio Eng. No. 7, 1954.
13. I. I. Perel'man, Control by Adaptive Programming. Avtomatika i Telemekhanika 19, 9 (1958).
14. V. I. Romanovskii, Discrete Markov Chains [in Russian]. Gostekhizdat, 1949.
15. V. S. Pugachev, The Theory of Random Functions and Applications to Automatic Control Problems [in Russian] Fizmatgiz, 1960.

THE QUANTIZATION ERROR IN AUTOMATIC CONTROL

S. M. Mandel'shtam

(Leningrad)

Translated from Avtomatika i Telemekhanika, Vol. 22, No. 6,
pp. 780-786, June, 1961
Original article submitted November 9, 1960

A method for determining the required parameter measurement frequency in discrete automatic control is considered. A series of assigned curves describing the variation of the controlled parameters and the assigned errors of the time quantization of the function serve as the initial calculation data. One of the possible methods for reducing the time quantization error is described.

In the practice of the automatic control of industrial processes, there are often cases where $f(t_0 + h_t i)$ values, where $i = 0, 1, 2, 3, \ldots$, which are arbitrarily taken at certain time intervals $h_t = \text{const}$, are measured or recorded instead of the total function variation in time. The advisability of such time discretization can be explained by the striving for a manifold utilization of individual elements in the apparatus or of the coupling channel in cyclic interrogation of the parameters as well as by a tendency to use digital recording.

The replacement of a continuous function by a time-series of instantaneous values leads to irreversible losses ΔI in the information value of the measurements, and, consequently, to an additional error δ_t [1]. With a greater frequency of readings, this error will correspondingly decrease. However, if the measurements are made too often, the possibilities of increasing the time-duty of the equipment are limited, and, for many element types, this leads to a reduction of their lifetime. On the other hand, if the readings are not taken often enough, the error δ_t attains inadmissible values. Hence the necessity of finding a mathematical relationship between the number of readings per unit time and the time quantization error.

In a number of papers concerned with discrete measurements (see, for instance, [3]), linear interpolation (Newton's formulas) is used for estimating this error. In this, only the mutual relationship between two adjacent values $f(t_i)$ and $f(t_i + h_t)$ of the function is taken into account. It is natural that such statement of the problem leads to higher interpolation error values, since the information contained in the totality of discrete values is insufficiently utilized.

For a better approximation of the actual function at the points between readings, it is necessary to take into account several neighbouring values of the function, and, if it is possible to follow the autocorrelation function over the entire interval in question, to take into account all points of the given interval.

A notable feature of a number of functions that characterize the variation of many technological parameters is their periodicial recurrence. It is understood that we do not consider ideal periodicity (the information obtained as a result of measurements would be otherwise equal to zero), but only a periodical repetition of the general character of changes in the function. The cause of this fact, which has been confirmed by the processing of a large number of curves, is the periodicity of many technological processes. The advisability of applying harmonic analysis to the curve describing changes in the technological process is an important consequence of this situation.

We shall now formulate the problem thus stated.

Let a certain arbitrary function be given in the time interval from t_0 to t_n. We shall assume that the chosen interval is sufficient for estimating the behavior of the function* and that the readings are taken during time in

* For a periodic function, the sufficient interval coincides with a single period. The determination of the period for actual curve does not involve difficulties if the technological cycle duration is given or if the presence of 24-hr (changeable) recurrence is known.

tervals that are negligibly small in comparison with the h_t intervals (Fig. 1). It is necessary to express the error incurred in approximating a graphically assigned curve by a sum of terms of a Fourier series with a finite number of terms. It is assumed that the function satisfies the Dirichlet conditions.

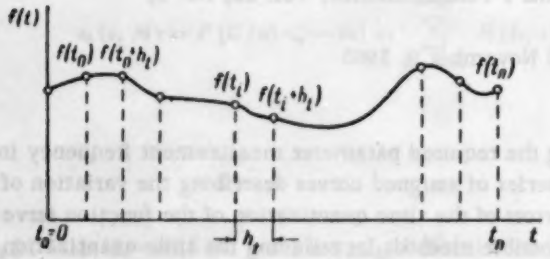


Fig. 1. Graph of the technological parameter in variation time.

Let us expand the function into a Fourier series with a number of terms that is generally limited by the highest reliably calculable finite difference of the parameter to be measured. It is natural that, in the case of such expansion, the function under consideration (strictly speaking, a nonperiodic function) will not be approximated by the obtained series beyond the boundary of the section under investigation; however, this is of no importance for subsequent considerations.

Let us designate by m the number of series terms obtained in this expansion. If we take the spectrum of the curve to be analyzed is limited by the frequency F_m, we have $\delta_m = 0$, where δ_m is the root-mean-square error in approximating a continuous curve by a sum of Fourier terms.

In this case of such expansion, any actual curve is divided into 12, 24, 36, or 48 intervals in dependence on the required accuracy and the behavior of the function.

In any case, it is not advisable to divide the curve into more than 64 to 72 intervals, since this would lead to very cumbersome calculations.

The coefficients are calculated by using the well-known Bessel functions:*

* The complexity of such analysis must not be exaggerated. Even in the case of division into 48 intervals, only the following trigonometric function values figure in the calculations:

$$\sin 0 = \cos\frac{\pi}{2} = 0;\ \sin\frac{\pi}{12} = \cos\frac{5\pi}{12} = 0,26;\ \sin\frac{\pi}{6} = \cos\frac{\pi}{3} = 0,50;$$

$$\sin\frac{\pi}{4} = \cos\frac{\pi}{4} = 0,71;\ \sin\frac{\pi}{3} = \cos\frac{\pi}{6} = 0,87;\ \sin\frac{5\pi}{12} = \cos\frac{\pi}{12} = 0,97;$$

$$\sin\frac{\pi}{2} = \cos 0 = 1.$$

All calculations can be readily performed by mechanical computers, without mentioning the possibility of using harmonic analyzers, electronic evolvers, etc. If special computing devices are not available, one can use grouping circuits (circuits designed by Nancy Klock or M. Serebrennikov in the case of 48 ordinates) or patterns (for instance, A. M. Lopshits or Zipperer patterns), or one can dispense with Bessel functions and use any of the analytical methods, with the exception, naturally, of methods for processing obliquely symmetrical periodic curves or other methods the derivation of which is connected with conditions unacceptable in the case under consideration. In other words, the practical harmonic analysis should be performed by using any of the possible methods with which the designer is familiar.

$$a_0 \approx \frac{2}{m_1} \sum_{k=0}^{m_1-1} f(t_k),$$

$$a_i \approx \frac{2}{m_1} \sum_{k=0}^{m_1-1} f(t_k) \cos \frac{2ki\pi}{m_1}, \qquad (1)$$

$$b_i \approx \frac{2}{m_1} \sum_{k=0}^{m_1-1} f(t_k) \sin \frac{2ki\pi}{m_1},$$

where $m_1 = 12k$ ($k = 1, 2, 3, \ldots$) is the number of intervals, and $i = 1, 2, \ldots, m_1/2$.

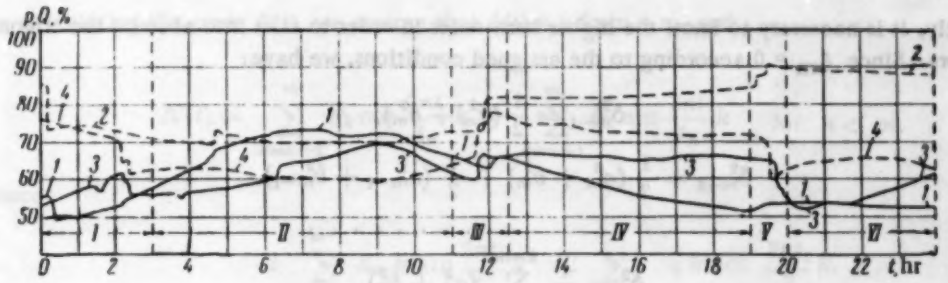


Fig. 2.

The approximating equal will be written as

$$S_{m_1}(t) = \frac{a_0}{2} + \sum_{k=1}^{k=m_1} a_k \cos \frac{2\pi t}{t_n} k + \sum_{k=1}^{k=m_1} b_k \sin \frac{2\pi t}{t_n} k.$$

If $m_1 > m$, then $a_{m+1} = a_{m+2} = \ldots = a_{m_1} = 0$ and $b_{m+1} = b_{m+2} = \ldots = b_{m_1} = 0$.

Therefore,

$$f(t) \approx S_m(t) = \frac{a_0}{2} + \sum_{k=1}^{m} a_k \cos \frac{2\pi t}{t_n} k + \sum_{k=1}^{m} b_k \sin \frac{2\pi t}{t_n} k. \qquad (2)$$

Nothing that the above method offers possibilities for the trigonometric interpolation of the initial function at the points between readings, we shall estimate the root-mean-square error incurred in approximating the function f(t) by a sum of n terms of the Fourier series for $n < m$.

By definition, we have

$$\delta_n^2 = \frac{1}{t_n} \int_0^{t_n} \Delta^2(t)\, dt = \frac{1}{t_n} \int_0^{t_n} \{f(t) - \frac{a_0}{2} = \sum_{k=1}^{n} a_k \cos \frac{2\pi t}{t_n} k + \sum_{k=1}^{k=m_1} b_k \sin \frac{2\pi t}{t_n} k\}^2\, dt.$$

It can be shown that

$$\delta_n^2 = \frac{1}{t_n}\int_0^{t_n} [f(t)]^2\,dt - \frac{a_0^2}{4} - \frac{1}{2}\sum_{k=1}^{n}(a_k^2 + b_k^2). \tag{3}$$

Nothing that the addition of a new term will not change the values of the coefficients considered, we shall write:

$$\delta_{n+1}^2 = \frac{1}{t_n}\int_0^{t_n} [f(t)]^2\,dt - \frac{a_0^2}{4} - \frac{1}{2}\sum_{k=1}^{n+1}(a_k^2 + b_k^2).$$

Hence, we conclude that the addition of a single new term to the n terms that have been considered reduces the error by $\frac{1}{2}(a_{n+1}^2 + b_{n+1}^2)$.

Consequently, it is necessary to know the higher harmonics in order to take into account the trigonometric interpolation errors. Since $\delta_m = 0$ according to the assigned conditions, we have:

$$\begin{aligned}
\delta_{m-1}^2 &= \frac{1}{2}(a_m^2 + b_m^2),\\
\delta_{m-2}^2 &= \frac{1}{2}(a_m^2 + b_m^2) + \frac{1}{2}(a_{m-1}^2 + b_{m-1}^2),\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
\delta_n^2 &= \frac{1}{2}\sum_{k=n+1}^{k-m}(a_k^2 + b_k^2).
\end{aligned} \tag{4}$$

By comparing expressions (3) and (4), we obtain:

$$\frac{2}{t_n}\int_0^{t_n} [f(t)]^2\,dt = \frac{a_0^2}{2} + \sum_{k=1}^{m}(a_k^2 + b_k^2).$$

The last relationship represents the closed-loop equation for a function with a bounded spectrum, which indicates that the above reasoning is correct.

Thus, in practice, if we have a curve that characterizes changes in the technological parameter which is being telemetered with the time quantization error δ_t, this curve should be decomposed by dividing it into m_1 sections. After we determine m + 1 coefficients that are different from zero, we successively calculate δ_{m-1}^2, δ_{m-2}^2, ... up to $\delta_{n+1}^2 \geqslant \delta_t^2 \geqslant \delta_n^2$ *. by using Eqs. (4). The latter determines in a well-defined manner the number of necessary readings during the time $t_n - t_0$ according to the formula

$$n_{req} = 2n.$$

In certain cases, the so-called average error $\sigma_{av} = 0.7979\ \delta$ of a single measurement is limited instead of the root-mean-square error.

This, naturally, does not change the analytical procedure, with the difference, however, that δ is calculated up to the value

$$\delta_{n+1}^2 \geqslant 2.04\,\sigma_{av}^2 \geqslant \delta_n^2.$$

* The equality must be understood in the sense of the accepted calculation accuracy, while the general accuracy estimate can be considered as the root-mean-square error $\delta_{oa} = \delta_{ind}/\sqrt{m_1}$, where δ_{ind} is the error of a single f(t) reading (i.e., the diagram recording error), and m_1 is the number of sections into which the function is divided.

It should be noted that, in some cases, a uniform approximation instead of the approximation in the sense of least squares is required.

We assume that the curve characterizing the time variation of the parameter to be measured does not have discontinuties (otherwise, the problem would not have a solution due to the Gibbs phenomenon). The continuity of the function makes it possible to state that such an approximation exists on the basis of the Weierstrass theorem. We shall try to obtain this approximation by means of a similar Fourier series. Since the $\sum_{k=0}^{k=m} a_k$ and $\sum_{k=1}^{m} b_k$ series are absolutely convergent, the series

$$\sum_{k=0}^{m} a_k \cos \frac{2\pi t}{t_n} k + \sum_{k=1}^{m} b_k \sin \frac{2\pi t}{t_n} k$$

are consequently uniformily convergent, and the Fourier series can provide a uniform approximation.

For calculating the error $\Delta(t)$, we assume that $\Delta(t) = 0$ for $n = m$, and

$$\Delta(t) = \sum_{k=n+1}^{m} a_k \cos \frac{2\pi t}{t_n} k + \sum_{k=n+1}^{m} b_k \sin \frac{2\pi t}{t_n} k \quad \text{for } n < m.$$

Hence,

$$2 \sum_{k=n+1}^{m} a_k k \sin \frac{2\pi t_i}{t_n} k = 2 \sum_{k=n+1}^{m} {}_k k \cos \frac{2\pi t_i}{2} k,$$

where t_i is a certain point in the interval under consideration ($t_0 \leq t_1 \leq t_n$), which corresponds to the maximum value of $\Delta(t)$.

It can be readily deduced that, for $n = m-1$,

$$t = \frac{t_n}{2\pi m} \operatorname{arc\,tg} \frac{b_m}{a_m}$$

and

$$\Delta_{max1} = b_m \sin \operatorname{arc\,tg} \frac{b_m}{a_m} + a_m \cos \operatorname{arc\,tg} \frac{b_m}{a_m};$$

for

$$\Delta_{max2} \approx \sum_{k=m-2}^{m} a_k \cos \operatorname{arc\,tg} \frac{b_m + b_{m-1}}{a_m + a_{m-1}} + \sum_{k=m-2}^{m} b_k \sin \operatorname{arc\,tg} \frac{b_m + b_{m-1}}{a_m + a_{m-1}}.$$

Further calculations lead to a system of transcendental equations, which can be reduced to high-power algebraic equations by means of the de Moivre formulas. However, the solution of such equations with respect to t_i is very difficult, and results which would be suitable for practical calculations have not yet been obtained.

The case of the coincidence of ordinates with the greatest error with respect to the time axis is of definite practical interest. It is natural that the maximum possible time discretization error occurs in this case. The condition for the coincidence of ordinates can be written as

$$\frac{b_m}{a_m} = \frac{mb_m + (m-1)\, b_{m-1}}{ma_m + (m-1)\, a_{m-1}}$$

or in a different form: $\frac{a_{m-1}}{b_{m-1}} = \frac{a_m}{b_m} = \ldots$

In this, $\Delta_{\max \Sigma} = \Delta_{\max 1} + \Delta_{\max 2} + \ldots$, where

$$\Delta_{\max i} = \sum_{k=m-i}^{m} a_k \cos \operatorname{arc\,tg} \frac{b_m}{a_m} + \sum_{k=m-i}^{m} b_k \sin \operatorname{arc\,tg} \frac{b_m}{a_m}.$$

In all other cases, it is apparently more convenient to use harmonic analyzers. Thus, for instance, the harmonic analyzer designed by Michelson and Straton [2] provides 80 expansion terms and makes it possible to synthesize a function from an arbitrary number of these terms by observing the pattern of successive approximations.

After processing one technological parameter curve, the designer can greatly simplify this calculations in the remaining cases by using the following procedure. In the case where the number of readings is found with respect to the assigned root-mean-square error, the Fourier coefficients can be calculated only in the expected region of harmonics with numbers from i = n to i = m. In the case of calculations with respect to the maximum absolute error, which can be reduced to the solution of a series of transcendental equations, it is most convenient to use the relaxation method assuming the first approximation from the calculations of the previous curves.

As a result of processing a large number of curves, we obtain functions that characterize the probability distribution of the errors under consideration for different interrogation frequencies, which makes it possible to obtain the average measurement error.

As far as it is known to the author, devices with Δt = const. are used almost exclusively in modern systems with cyclic interrogation. However, it is advisable to design a rational system in a different way.

In most cases, there is an excess of information, which is connected with the statistical distribution of the derivative of the function to be measured in different time intervals. By taking into account this information, it is possible to reduce δ_t without a superfluous increase in the parameter interrogation frequency. We shall clarify this by considering figure 2.

This figure shows the curves of relative gas discharge values Q (curves 2 and 4) and pressure values p (curves 1 and 3) at the Leningrad gas distribution station over a period of two days in winter time. It can be shown that the discharge and pressure curves on other winter days will lie in a very narrow region near the above curves while preserving approximately the same shape. Only the curves obtained on holidays will be an exception to this.

Let us divide the entire interval under consideration into sections I to VI. We notice that the function variation rate is different for each section. If we calculate the maximum interpolation error in each of these sections for an increment h_t = const, we can readily see that they are different in each section. Hence, it is convenient to chose a separate h_t value for each section. For instance, for δ_t = 1.5%, we have: h_{t_1} = 10 min; h_{t_2} = 1 hr; h_{t_3} = = 10 min; h_{t_4} = 2 hr; h_{t_5} = 14 min; h_{t_6} = 50 min.

The curves shown in figure 2 pertain to normal operating conditions. Under emergency conditions, it is apparently advisable to impose on the system the regime of cyclic interrogation of all units or of a certain parameter group with the maximum speed. Such a change in h_t can be brought about by the emergency signal, which is in many cases provided by means of a special data transmitter.

SUMMARY

1. Besides errors inherent in instruments and measurement methods, an additional and unavoidable time quantization error arises in automatic control processes.

The root-mean-square value δ_t of this error can be calculated for a given function with a bounded spectrum by using trigonometric interpolation according to the formula

$$\delta_t = \sqrt{\frac{1}{2} \sum_{i=n+1}^{m} (a_i^2 + b_i^2)},$$

where a_i and b_i are the coefficients obtained by expanding the given function into a Fourier series, and 2n is the number of discrete readings.

2. The above time quantization error for continuous functions can readily be calculated by means of harmonic analyzers and synthesizers.

3. For periodic functions, the time quantization error can be reduced by a suitable division of the function into regions with a specific interrogation interval h_t for each region.

LITERATURE CITED

1. L. Brillouin, Science and Information Theory [Russian translation] (Fizmatgiz, 1960).
2. V. L. Goncharov, Theory of the Interpolation and Approximation of Functions [in Russian] (Gostekhizdat, 1954).
3. V. N. Khlistunov, "On the approximation error in using discrete measurement methods," Priborostroenie No. 5, 1960.

THEORY OF A TWO-CORE MAGNETIC AMPLIFIER WITH RECTIFICATION OF THE EVEN HARMONICS BY MEANS OF A NONLINEAR RESISTANCE

Dai Tse-Hsin

(Moscow)

Translated from Avtomatika i Telemekhanika, Vol. 22, No. 6,
pp. 787-794, June, 1961
Original article submitted October 22, 1960

The paper analyzes the operation of a magnetic amplifier network based on the use of even voltage harmonics that appear across the terminals of the output windings of the simplest two-core magnetic amplifier when a dc magnetizing field is present.

The author analyzes the operation of a magnetic amplifier network based on two nonidentical cores with an output winding connected in series with a symmetrical nonlinear resistance.

Magnetic amplifiers based on the use of the even voltage harmonics which appear across the terminals of the control or output windings can be subdivided into two groups [1]. The first group includes magnetic voltage amplifiers with a frequency-doubled output. They have a high sensitivity but feature a cumbersome construction and involve complex auxiliary equipment, in view of the necessity of connecting high-quality multiple-section band pass filters at the output of the amplifier and a blocking filter in series with the excitation windings. One of the basic reasons for the appearance of a parasitic ac voltage at the amplifier output is the variation of the characteristics and the violation of the identical nature of the cores; this requires the cores to be selected in the most thorough way possible.

The second group includes magnetic amplifiers which are excited by means of a large ac voltage. Pulses whose amplitude and duration are functions of the magnetizing signal appear at the terminals of the output winding of such an amplifier (across the terminals K_1-K_2 in the circuit shown in Fig. 1). By selective rectification of these pulses it is possible to obtain a dc current component at the output of the magnetic amplifier when a signal is present at its input; this dc component can be applied to the next stage of a multistage magnetic amplifier. The pulses can be rectified by means of a nonlinear resistance R_{ne} with a symmetrical volt-ampere characteristic. Ideal for this purpose is a nonlinear resistance which has a volt-ampere characteristic such as is shown in Fig. 2.

In this paper we assume that the symmetrical nonlinear resistance has an ideal volt-ampere characteristic such as that shown in Fig 2. If we form the symmetrical nonlinear resistor R_{ne} by connecting a silicon stabilitron of the type "D808-813" into the diagonal of a rectifier bridge consisting of semiconductor diodes (cf. Fig. 3), then

its characteristic will be close to ideal. Magnetic amplifiers of the second group are distinguished by their simple circuity when a symmetrical nonlinear resistor is used to rectify the even voltage harmonics, and do not require filters for suppressing the odd harmonics in the output circuit. An experimental investigation shows that magnetic amplifiers of the type shown in Fig. 1 have a high null stability for changes in the supply voltage and frequency, or for temperature changes over wide limits [2]. However, the literature provides no analysis of the causes for the null drift and does not explain the physics governing the operation of the magnetic amplifier proper; quantitative relationships between the basic parameters of the magnetic amplifier circuit are also absent in the literature.

The purpose of this paper is to analyze the operation of a magnetic amplifier of the type shown in Fig. 1 and to derive quantitative relationships between the basic parameters of the magnetic amplifier.

Below, we shall demonstrate that the output current reaches its maximum when the control current I_c is equal to the current value I_s for which the core is saturated. Therefore, if we approximate the magnetization curves in the manner shown in Fig. 4 we shall lose the portion of the amplifier characteristic corresponding to the values $I_c < I_s$; this is undesirable.

We shall approximate the magnetization curve by means of three rectilinear segements as shown in Figs. 5-9. In that case,

$$\mu = \frac{dB}{dH} = \text{const} \quad \text{for} \quad |I| < |I_s|,$$

$$\mu = \frac{dB}{dH} = 0 \quad \text{for} \quad |I| > |I_s|.$$

The paper cites the results of an analysis for the case where the cores differ somewhat. The case of identical cores is a particular case of the more general case studied here. For the circuit of Fig. 1 the presence of a small difference between the cores is not only not harmful but is even useful from the point of view of sensitivity (1); this follows since in the case of slightly differing cores initial pulses exist across the terminals of the output winding when there is no signal, and these pulses keep the symmetrical nonlinear resistor "open." The refore for the circuit in Fig. 1 we do not require the rigorous identity of cores which is required by magnetic voltage amplifiers that operate according to the frequency doubling principle.

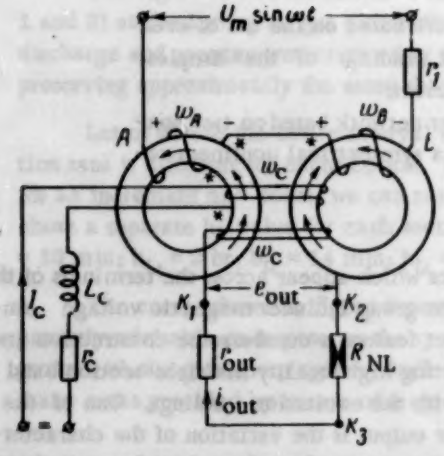


Fig. 1. The circuit diagram for the magnetic amplifier.

We assume that nonidentical cores are defined as cores in which the inductances of identical windings, and (or) the saturation fluxes, and (or) the current for which the cores saturate are not identical. $\Phi_{sA} \neq \Phi_{sB}$, $L_A \neq L_B$, $I_{sA} \neq I_{sB}$. In order to be specific we shall study a sufficiently general case for which

$$\Phi_{sB} > \Phi_{sA} \quad \text{and} \quad L_B < L_A. \tag{1}$$

Here $I_{sB} > I_{sA}$, since

$$L_A = w_A \frac{\Phi_{sA}}{I_{sA}} 10^{-8} \quad \text{and} \quad L_B = w_B \frac{\Phi_{sB}}{I_{sB}} 10^{-8}.$$

For simplicity in notation we shall assume that all the windings have an identical number of turns:

$$w_A = w_B = w_C = w_{out} = w. \tag{2}$$

Then

$$I_{sA} = w \frac{\Phi_{sA}}{L_B} 10^{-8} \quad \text{and} \quad I_{sB} = w \frac{\Phi_{sB}}{L_B} 10^{-8}. \tag{3}$$

The operation of the magnetic amplifier shall be studied for the case of very weak signals when

$$I_c \ll \frac{U_{av}}{r_1},$$

where r_1 is the resistance in the excitation circuit, and U_{av} is the average value of the supply voltage.

It should be noted that in ordinary magnetic amplifiers we are interested in the current through the supply circuit (which is the load circuit) as a function of the control current. For the magnetic amplifier under study in this paper we are interested in the current through the output winding, which is now the load circuit, rather than the current in the excitation circuit. Since a magnetic amplifier of the type shown in Fig. 1 is designed for amplifying very weak signals, it follows that the current in the excitation circuit is practically independent of the control current. In order to avoid any effect of hysteresis on the null drift the magnetic amplifier shown in Fig. 1 is in practice excited using a large ac voltage; thus, during a portion of each half-cycle both cores are in a saturated state. Under these conditions the magnitude of the current in the excitation circuit is limited by the resistance of that circuit.

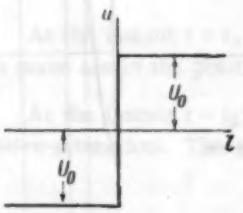


Fig. 2. Idealized characteristics for the nonlinear resistance.

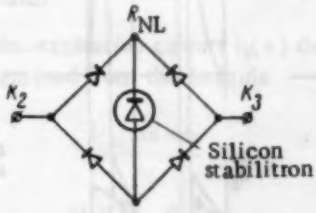


Fig. 3.

Fig. 4.

We shall first study the operation of the amplifier when there is no signal (Fig. 5). For $t < t_1$ both cores are in a state of negative saturation. At the instant $t = t_1$ when the negative exciting current $i_1(-)$ decreases to the value $-I_{sB}$ in absolute magnitude (the point 1' on the magnetization curve for core B in Fig. 5h), the core B emerges from the saturation state. Under these conditions the core A is still saturated (point 1 on the magnetization curve for core A in Fig. 5g). The instant t_1 is determined (when the number of turns is equal) by the condition

$$\frac{U_m \sin \omega t_1}{r_1} = -I_{sB}. \tag{4}$$

For small resistance r_1, $|r_1 I_{sB}| \ll U_m$ and $t_1 \approx 0$; therefore in our case the current I_{sB} for good permalloy is very small when the dimensions of the cores and the number of turns on the excitation windings are taken into account.

For $t_1 \leq t \leq t_2$ only the flux through core B varies; an emf is induced in the output winding, and this emf has a polarity opposite to the polarity of the exciting voltage (Fig. 5e).

At the instant $t = t_2$ at which the current in the excitation circuit reaches the value $i_1(-) = -I_{sA}$ the states of the cores A and B are characterized by the points 2 and 2', respectively (Figs. 5g and 5h). The core A emerges from the saturation state. The instant t_2 is determined (neglecting the voltage drop $r_1 i_1$) from the equation

$$\int_{t_1}^{t_2} U_m \sin \omega t \, dt \approx -w 10^{-8} (\Phi_{1'} - \Phi_{2'}) = w 10^{-8} \Phi_{sB} - L_B I_{sA}. \tag{5}$$

For $t_2 \leq t \leq t_3$ an emf $e_{out} = e_A - e_B$ is induced in the output winding; this emf is produced by the joint variations of the fluxes Φ_A and Φ_B. The emf $e_{out} = e_A - e_B$ is not equal to zero, since the inductance of core A is

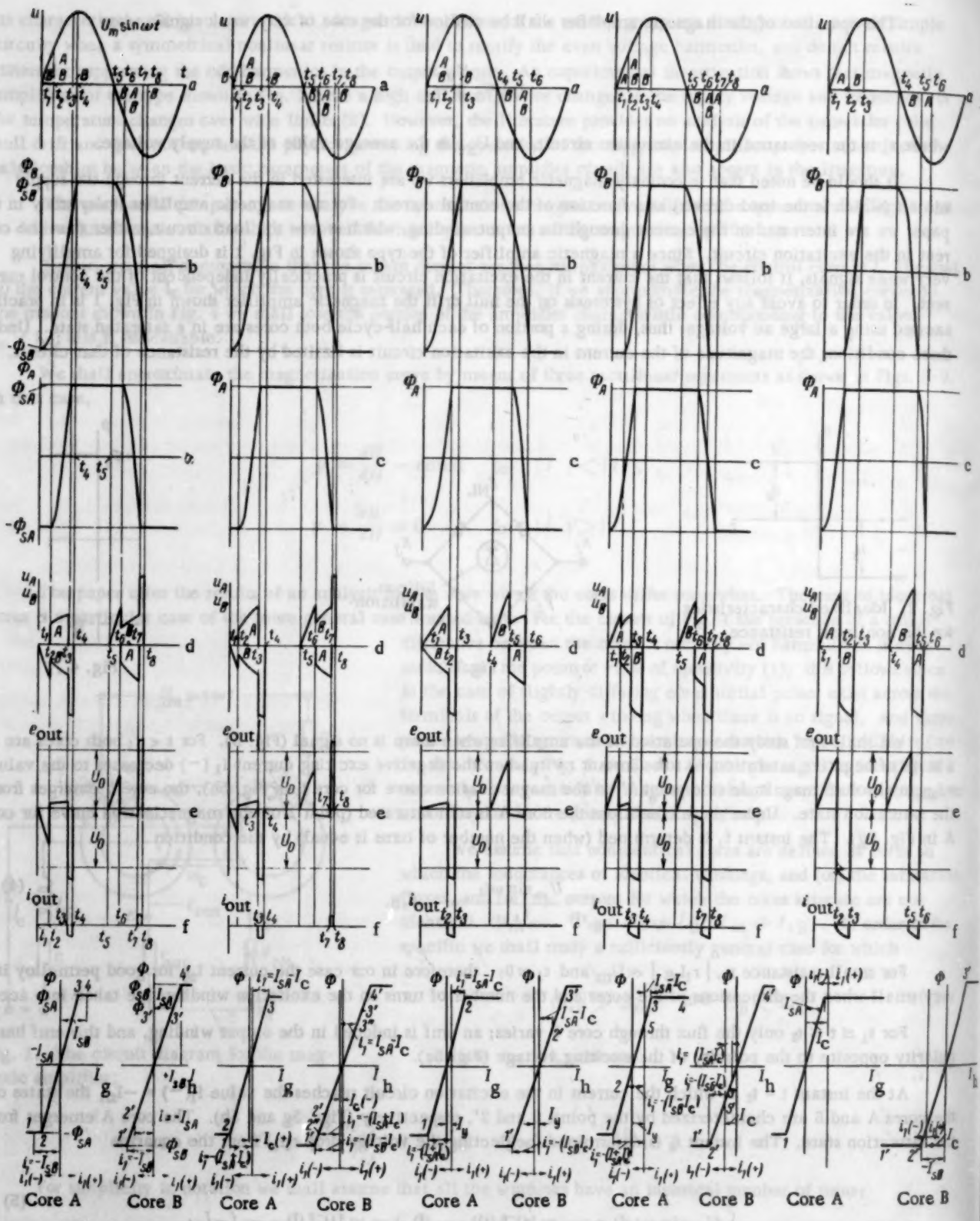


Fig. 5 Fig. 6 Fig. 7 Fig. 8 Fig. 9

Fig. 5. Case of $I_C = 0$

Fig. 6. Case of $0 < I_C < \frac{I_{sB} - I_{sA}}{2}$

Fig. 7. Case of $I_C = \frac{I_{sB} - I_{sA}}{2}$

Fig. 8. Case of $\frac{I_{sB} - I_{sA}}{2} < I_C < \frac{I_{sB} + I_{sA}}{2}$

Fig. 9. Case of $I_C = \frac{I_{sB} + I_{sA}}{2}$

greater than the inductance of core B ($L_A > L_B$). The emf e_{out} has the same polarity as the voltage across the terminals of the excitation windings for the core A.

At the instant $t = t_3$ the stages of the cores A and B are characterized by the points 3 and 3' on the corresponding magnetization curves. The core A saturates. The instant t_3 is determined by the formula

$$\int_{t_1}^{t_3} U_m \sin \omega t \, dt \approx w 10^{-8} (2\Phi_{sA} + \Phi_{sB} + \Phi_{3'}) = w 10^{-8} (2\Phi_{sA} + \Phi_{sB}) + L_B I_{sA}. \quad (6)$$

For $t_3 \le t \le t_4$ only the flux Φ_B changes. A pulse appears at the terminals of the output windings, since in that time interval the value of the exciting voltage is large and this value is transferred to the output winding by means of the variation of the flux Φ_B. The instant t_4 is determined from the equation

$$\int_{t_1}^{t_4} U_m \sin \omega t \, td \approx w 10^{-8} (2\Phi_{sA} + 2\Phi_{s_B}). \quad (7)$$

At the instant $t = t_4$ the state of the core B is expressed by the point 4' and the core B saturates. For $t_4 \le t \le t_5$ both cores are in the positive saturation state.

At the instant $t = t_5$ the current in the excitation circuit $i_1(+)$ decreases to $+I_{sB}$, and core B emerges from positive saturation. The instant t_5 is determined from the formula

$$\frac{U_m \sin \omega t_5}{r_1} = I_{sB}. \quad (8)$$

From the instant $t = t_5$ the fluxes Φ_B and Φ_A begin to vary in reverse order.

Assume for simplicity that the magnitude of the emf e_{out} induced in the output winding is greater than the threshold voltage U_0 of the nonlinear resistance only in the case when e_{out} is in the form of a pulse. Then we obtain the following formula for the current in the output circuit:

$$i_{out} = \frac{e_{out} - U_0}{r_{out} + r_1 \left(\frac{w_{out}}{w_1} \right)^2} = \frac{e_{out} - U_0}{r}, \quad (9)$$

where

$$r = r_{out} + r_1 \left(\frac{w_{out}}{w_1} \right)^2 = r_{out} + r_1.$$

when the equality of the turns w_{out} and w_1 is taken into account.

When there is no signal the average value of the current is

$$\frac{1}{T} \int_0^T i_{out} \, dt = 0.$$

The mmf produced by the control current I_c and the excitation current i_1 add in core A and subtract in core B. When the control current I_c shown in Figs. 6g and 6h is present, only the flux Φ_B varies at first from $-\Phi_{sB}$ (1') to the point 2" on the magnetization curve.

The values of the current i_1 at these points will be :

for the point 1',

$$i_1(-) = -(I_{sB} - I_c),$$

for the point 2"

$$i_1(-) = -(I_{sA} + I_c).$$

The point 2" for the core B corresponds to a value of the current i_1 for which the core A emerges from the negative saturation state (point 2). After that the joint variation of the fluxes Φ_A and Φ_B begins. The flux Φ_B varies from point 2" and 3", and the flux Φ_A varies from the point 2 to the value $+\Phi_{sA}$ (point 3). After that only Φ_B varies from the point 3" to $+\Phi_{sB}$ (the point 4').

A variation of the flux Φ_B from the point 3" to the point 4' induces a pulse in the output winding. A reverse variation of the flux Φ_B from the point 2" to the point 1' during the next negative half cycle also induces a pulse. Since the flux increment $(\Phi_{4'} - \Phi_{3''}) > (\Phi_{4'} - \Phi_{3'})$ and $(\Phi_{1'} - \Phi_{2''}) < (\Phi_{1'} - \Phi_{2'})$, it follows that the duration of the pulse for $t_3 \leq t \leq t_4$ becomes greater when the signal is present while the duration of the pulse for $t_7 \leq t \leq t_8$ becomes less (Fig. 6e).

We shall determine the value of control current for which the duration of the pulse goes to zero for $t_7 \leq t \leq t_8$ (Figs. 5e and 6e). It is evident that for the equation (Fig. 7g and 7h)

$$I_{sA} + I_C = I_{sB} - I_C,$$

which corresponds to the formula

$$I_C = \frac{I_{sB} - I_{sA}}{2}, \tag{10}$$

the flux Φ_A and Φ_B begin to change simultaneously: Φ_A changes from $-\Phi_{sA}$, and Φ_B changes from $-\Phi_{sB}$. After the equation $\Phi_A = +\Phi_{sA}$ is reached only the flux Φ_B varies from point 2' to the point 3' (Fig. 7h), and a pulse is induced in the output winding. During the next negative half-cycle the fluxes Φ_A and Φ_B vary in reverse order; first only the flux Φ_B varies from the point 3' to the point 2'. After that the flux Φ_A and Φ_B vary jointly and simultaneously reach $-\Phi_{sA}$ and $-\Phi_{sB}$. Thus, during the negative half-cycle no pulse is induced in the output windings.

For $^1/_2 (I_{sB} - I_{sA}) \leq I_C \leq {}^1/_2 (I_{sB} + I_{sA})$ two pulses of emf appear with only negative polarity (Fig. 8e). We shall demonstrate that for $I_C = {}^1/_2(I_{sB} - I_{sA})$ the output current reaches a maximum (Fig. 9).

Whereas at first only the flux Φ_A varies from $-\Phi_{sA}$ to $+\Phi_{sA}$ and after that the flux Φ_B varies from $-\Phi_{sB}$ to $+\Phi_{sB}$, it follows that the output current i_{out} reaches a maximum as shown in Fig. 9. Under these conditions the magnitude of the current i_1 determined from the ampere-turns equation for the point $+\Phi_{sA}$ is equal to the magnitude of the current i_1 determined from the ampere-turns equation for the point $-\Phi_{sB}$ (Figs. 9g and 9h):

For the point $+\Phi_{sA}$

$$i_1(-) = I_{sA} - I_C;$$

For the point $-\Phi_{sB}$

$$i_1(-) = -I_{sB} + I_C,$$

whence

$$I_C = \frac{I_{sB} + I_{sA}}{2}. \tag{11}$$

After the control current I_C reaches the value $^1/_2(I_{sB} + I_{sA})$, a further increase in I_C cannot cause an increase in the dc component of the output current i_{out}. In other words, for $I_C = {}^1/_2(I_{sB} + I_{sA})$ the dc component of the output current reaches a maximum.

We shall derive the relationships between the dc component I_{av} of the output current and the signal I_C which are of interest to us for the case of a weak signal $I_C \ll {}^1/_2(I_{sB} + I_{sA})$; here it is possible to assume that the time interval during which the output voltage pulse exists is sufficiently small. Then we can assume that the emf e_{out} is constant during the interval over which the pulse exists. The duration Δt of the pulse depends on how much the flux $\Delta\Phi$ varies:

$$U_m \Delta t = w 10^{-8} \Delta\Phi. \tag{12}$$

For the average values I_{pos} and I_{neg} of the output current during the positive and negative half cycles, respectively, we have

$$I_{pos} = \frac{2}{T} \frac{U_m - U_0}{r} \Delta t_{pos} \tag{13}$$

$$I_{neg} = \frac{2}{T} \frac{U_m - U_0}{r} \Delta t_{neg} \tag{14}$$

Substituting (12) into (13) and (14), we obtain:

$$I_{pos} = \frac{2}{T} \frac{1 - \frac{U_0}{U_m}}{r} w 10^{-8} \Delta\Phi_{pos} \tag{15}$$

$$I_{neg} = \frac{2}{T} \frac{1 - \frac{U_0}{U_m}}{r} w 10^{-8} \Delta\Phi_{neg} \tag{16}$$

The dc component of the output current is equal to

$$I_{av} = I_{pos} - I_{neg} = \frac{2}{T} \frac{1 - \frac{U_0}{U_m}}{r} w 10^{-8} (\Delta\Phi_{pos} + \Delta\Phi_{neg}). \tag{17}$$

First we shall study the relationships between the control currents and the dc component of current in the output winding for

$$0 \leqslant I_c \leqslant {}^1/_2 (I_{sB} - I_{sA}).$$

As demonstrated in Fig. 5, for $I_c = 0$ the flux variation is $-\Delta\Phi_{neg} = \Phi_4' - \Phi_3'$, $-\Delta\Phi_{pos} = \Phi_1' - \Phi_2'$, and $\Delta\Phi_{pos} + \Delta\Phi_{neg} = 0$; therefore, on the basis of Eq. (17) the dc component of the current i_{out} is equal to zero.

When the signal I_c is present I_c $(0 < I_c < {}^1/_2 (I_{sB} - I_{sA}))$ the flux Φ_B varies by the amount $\Phi_1' - \Phi_2''$ instead of by the amount $\Phi_1' - \Phi_2'$ during the positive half-cycle (Fig. 6h); during the negative half-cycle it varies by the amount $\Phi_4' - \Phi_3''$ instead of by the amount $\Phi_4' - \Phi_3'$. The flux increments will be

$$-\Delta\Phi_{pos} = (\Phi_1' - \Phi_2'') - (\Phi_1' - \Phi_2') = -\Phi_2'' + \Phi_2', \tag{18}$$

$$-\Delta\Phi_{neg} = (\Phi_4' - \Phi_3'') - (\Phi_4' - \Phi_3') = -\Phi_3'' + \Phi_3'. \tag{19}$$

The flux increment over the entire cycle will be equal to

$$-\Delta\Phi_{pos} - \Delta\Phi_{neg} = -(\Phi_2'' - \Phi_2') - (-\Phi_3'' + \Phi_3') = -2(\Phi_2'' - \Phi_2'). \tag{20}$$

due to the presence of the signals.

From Figs. 6g and 5h it is evident that

$$-2(\Phi_2'' - \Phi_2') w 10^{-8} = 2[(I_{sA} + I_c + I_c) L_B - I_{sA} L_B] = 4 \frac{\Phi_{sB}}{I_{sB}} I_c w 10^{-8}. \tag{21}$$

Substituting (21) into (17), we obtain

$$I_{av} = -\frac{2}{T} \frac{1 - \frac{U_0}{U_m}}{r} w 10^{-8} \left(4 \frac{\Phi_{sB}}{I_{sB}}\right) I_c \quad \text{for} \quad 0 \leqslant I_c \leqslant \frac{I_{sB} - I_{sA}}{2}. \tag{22}$$

From formula (22) it is evident that the dc component of the output current i_{out} for $0 \leqslant I_c \leqslant {}^1/_2 (I_{sB} - I_{sA})$ is proportional to the signal I_c.

We shall study the relationships between the control current and the dc component of current in the output winding for

$$ {}^1/_2 (I_{sB} - I_{sA}) \leqslant I_c \leqslant {}^1/_2 (I_{sB} + I_{sA}). $$

During the positive half-cycle the flux Φ_B varies by the amount $\Phi_{4'} - \Phi_{3'}$, and during the negative half-cycle the flux Φ_A varies by the amount $\Phi_1 - \Phi_2$.

From the magnetization curves in Figs. 8g and 8h we determine flux increments:

$$ \begin{aligned} -\Delta\Phi_{heg}\, w\, 10^{-8} &= (\Phi_{4'} - \Phi_{3'})\, w\, 10^{-8} = [(\Phi_{4'} - \Phi_{5'}) - (\Phi_{3'} - \Phi_{5'})]\, w\, 10^{-8} = \\ &= [(I_c + I_{sB})\, L_B - (I_{sA} - I_c)\, L_B] = L_B\, [2I_c + (I_{sB} - I_{sA})] \end{aligned} \tag{23} $$

and

$$ \begin{aligned} -\Delta\Phi_{pos}\, w\, 10^{-8} &= -(\Phi_1 - \Phi_2)\, w\, 10^{-8} = -[(\Phi_1 - \Phi_5) - (\Phi_5 - \Phi_2)]\, w\, 10^{-8} = \\ &= [(I_{sA} + I_c)\, L_A - (I_{sB} - I_c)]\, L_A = L_A\, [2I_c - (I_{sB} - I_{sA})]. \end{aligned} \tag{24} $$

From (23) and (24) we find that the flux increment over the entire cycle is equal to

$$ (\Delta\Phi_{pos} + \Delta\Phi_{heg})\, w\, 10^{-8} = -L_A\, [2I_c - (I_{sB} - I_{sA})] - L_B\, [2I_c + (I_{sB} - I_{sA})]. \tag{25} $$

It should be noted that by substituting $I_c = {}^1/_2 (I_{sB} - I_{sA})$ into (21) or (25) we obtain a flux increment equal to $2\,\frac{\Phi_{sB}}{I_{sB}}\,(I_{sB} - I_{sA})$. in both cases.

Substituting (25) into (17), we obtain the expression for the dc component of the output current:

$$ I_{av} = -\frac{2}{T}\,\frac{1 - \frac{U_0}{U_m}}{r}\,[(L_B - L_A)\,(I_{sB} - I_{sA}) + 2\,(L_B + L_A)\,I_c]. \tag{26} $$

The dc component I_{av} is proportional to the control current I_c.

Equation (26) will not be valid if the inequality $I_c \leqslant {}^1/_2 (I_{sB} + I_{sA})$ is not valid, since under these conditions the magnitude of the pulse can no longer be considered constant. But such an operating mode of the magnetic amplifier is of no interest to us.

For the case of identical cores we obtain

$$ I_{av} = -\frac{2}{T}\,\frac{1 - \frac{U_0}{U_m}}{r}\, w \cdot 10^{-8} \cdot 4\,\frac{\Phi_s}{I_s}\, I_c. \tag{27} $$

from (22) or (26).

From (22), (26) and (27) it is evident that a slight difference between the cores ($L_B \approx L_A$, $I_{sB} \approx I_{sA}$) has only a slight effect on the current gain $K_I = I_{av}/I_c$.

SUMMARY

The analysis which we have performed for a two-core magnetic amplifier shows that for the case of non-identical cores initial pulses appear across the terminals of the output windings when there is no signal. When the signal is applied the pulses decrease in one direction and increase in the other direction.

We have derived the basic relationships between the dc component I_{av} of the output current, the control current I_c and the magnetic amplifier parameters.

We have demonstrated that a slight difference between the cores has very little effect on the current gain, and therefore no special selection of cores is required.

For the magnetic amplifier circuit shown in Fig. 1,a small difference in the cores (and therefore a variation in the core characteristics) has no effect on the amplifier sensitivity. The lower sensitivity threshold of the magnetic amplifier is limited to a considerable extent by the instability of the symmetrical nonlinear resistance (this is studied in [3]).

The facts that pulse voltages arise across the terminals of the output winding of the amplifier and that the input-output characteristic is linear within certain definite limits are verified experimentally.

The author expresses his appreciation to M. A. Rozenblat for his valuable remarks and advice.

LITERATURE CITED

1. M. A. Rozenblat, Magnetic Amplifiers [in Russian] (Soviet Radio Press, 1960).
2. B. A. Mazzeo, A Low-Level, High-Accuracy, D-C Magnetic Amplifier. Electrical Manufacturing Vol. 62, No. 5, November, 1958.
3. Dai Tse-Hsin, "The use of semiconductor diodes in the design of a symmetrical nonlinear resistance," Avtomatika i Telemekhanika 22, 3 (1961).

THE DISTRIBUTION OF THE PEAK DURATIONS FOR PULSE NOISE AT THE OUTPUT OF A REMOTE CONTROL UNIT

L. B. Venchkovskii

(Moscow)

Translated from Avtomatika i Telemekhanika, Vol. 22, No. 6, pp. 795-800, June, 1961
Original article submitted December 9, 1960

A study is made of the effect of pulse noise with a logarithmically normal amplitude distribution on a remote control unit. The probability density distribution is determined for the duration of the peaks at the output of the threshold unit for a low-pass filter of the RC type and for a filter with a frequency response corresponding to a Gaussian curve.

INTRODUCTION

The investigation of noise in a low-voltage power line rated at 0.4/6 kv [1], as well as other papers on the investigation of noise which is of industrial origin (cf. for example, [2]), have already made it possible to conclude that a rather broad class of industrial noise in a passband of the order of several kilocycles is of a pulse nature with a logarithmically normal amplitude distribution law. Therefore it is of interest to study the effect of this type of noise on remote control units which operate in a wide frequency band over a power line and over a radio channel.

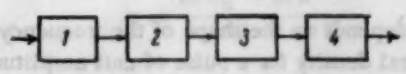


Fig. 1.

In order to estimate the noise immunity for a number of remote control systems along with the amplitude distribution it is also necessary to make use of the probability distribution density for the duration of the peaks at the output of the receiver. A somewhat simpler problem arises in finding the magnitude of the error caused by the noise in pulsed telemetering systems when it is possible to limit ourselves to determining the probability that the peak duration exceeds a specified value; in a number of cases this problem reduces to finding the average duration of the pulses at a specified level.

It is of interest to determine the probability distribution density for the duration of the pulse noise peaks at the output of a remote control unit (Fig. 1) consisting of a wide-band input filter 1, a linear inertialess amplitude detected a, a low-pass filter 3, and a threshold device 4. We shall perform our analysis for the case where the input filter has a frequency response such that when its input is subjected to a brief noise pulse, one pulse whose shape is determined entirely by the frequency response of the filter appears at its output.

Since the distribution of the amplitude envelope for the noise is usually measured at the output of the selective metering device, it is natural to assume that the logarithmically normal amplitude distribution of the pulses is determined at the output of the filter 1 and can be written as

$$p(\ln x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\ln x - \ln a)^2}{2\sigma^2}\right] \quad (0 \leqslant x < \infty), \tag{1}$$

where ln a and σ respectively represent the average and mean-square values of the normal distribution for ln x.

We shall also assume that the speed of response for the system is such that the noise signals at the output of the receiver remain of a pulse nature (i.e., superposition of their transient responses does not occur). Such an assumption applies to multichannel and fast-response remote control systems when the mean repetition frequency of the pulses is sufficiently small. For simplicity we shall assume that the input filter has a band so much wider than the output filter that the noise pulse spectrum remains constant within the limits of the passband for the filter 3. This does not reduce the general nature of the results, since they may be extended to include the case of commensurate pass bands for both filters when the mathematical computations are made somewhat more complex. For such sufficiently general assumptions the problem reduces to finding the distribution of the peak durations for a number of threshold voltage values when the filter 3 is subjected to pulses of identical shape governed by a logarithmically normal amplitude distribution law.

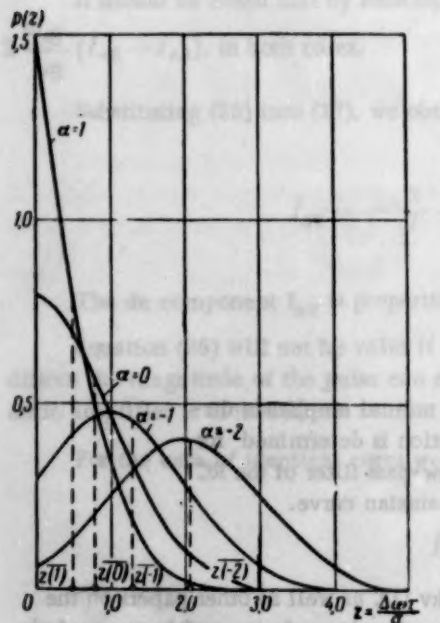


Fig. 2.

1. A Low-Pass Filter Consisting of a RC-Circuit

As we know, the frequency response for a RC low-pass filter can be written as

$$C(j\omega) = \frac{1}{1 + j\omega CR}, \tag{2}$$

where $\frac{1}{RC} = \Delta\omega$ is the passband width for the filter at the 0.7 level.

If the input of such a filter is subjected to a brief pulse of amplitude x whose spectral density is constant within the limits of the pass band $\Delta\omega$ and proportional to its amplitude x, then the shape of the pulse at the filter output can be determined from the formula

$$u_{out}(t) = C_0 \Delta\omega x e^{-\Delta\omega t}. \tag{3}$$

The proportionality coefficient C_0 depends on the shape of the frequency response of the input filter and represents the numerical value of the spectral density for a pulse of unit amplitude at the output of the filter 1 at its turned frequency ω_0.

The time τ during which the noise peak will exceed the voltage U_{th} corresponding to the operating threshold can be determined from the relationship

$$U_{th} = C_0 \Delta\omega x e^{-\Delta\omega\tau}, \tag{4}$$

whence

$$\Delta\omega\tau = \ln x - \ln U_n. \tag{5}$$

Here $U_n = U_{th}/C_0 \Delta\omega$ is the value of the threshold voltage normalized to the filter input.

It follows from formula (5) that the distribution for $\Delta\omega\tau$ coincides with the distribution for ln x with a shifted distribution center. Therefore for the probability density it is possible to write

$$p(\Delta\omega\tau) = \begin{cases} \dfrac{K_c}{\sqrt{2\pi}\sigma} \exp\left[-\dfrac{(\Delta\omega\tau - \ln a + \ln U_n)^2}{2\sigma^2}\right] & (0 \leqslant \Delta\omega\tau < \infty) \\ 0 & (\Delta\omega\tau < 0). \end{cases} \tag{6}$$

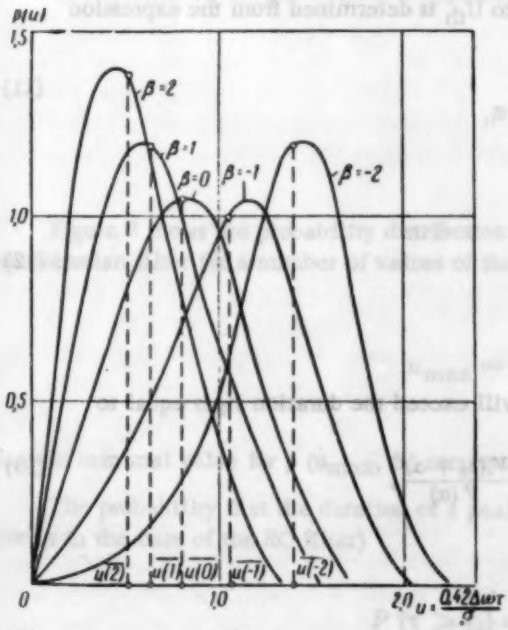


Fig. 3.

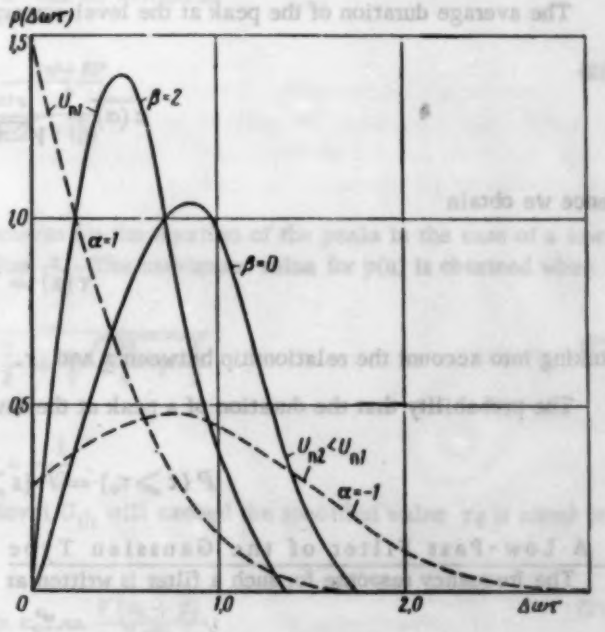


Fig. 4.

This formula represents the truncated normal distribution in which the coefficient K_c is determined from the condition governining the normalization of the probability density curve.

In order to achieve a graphical plot and to determine the numerical distribution parameters it is convenient to study the normalized distribution for $z = \Delta\omega\tau/\sigma$ and to express the magnitude of the threshold voltage in terms of the coefficient

$$\alpha = \frac{\ln U_n - \ln a}{\sigma}, \tag{7}$$

by representing it in the form

$$U_n = ae^{\alpha\alpha}. \tag{8}$$

As a result we can write

$$p(z) = \frac{e^{-\frac{(z+\alpha)^2}{2}}}{\sqrt{2\pi}V(\alpha)} \qquad (z \geqslant 0), \tag{9}$$

where

$$V(\alpha) = \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\infty} e^{-\frac{t^2}{2}} dt. \tag{10}$$

Figure 2 shows the family of distribution curves for the duration of the peaks at the output of an RC filter for a number of values of the parameter α. It is evident from the curves that as the cutoff voltage decreases, the probability distribution density for the duration of the peaks approaches a normal distribution with an average value close to $-\alpha$.

The average duration of the peak at the level corresponding to U_{th} is determined from the expression

$$\overline{z(\alpha)} = \frac{e^{-\frac{\alpha^2}{2}}}{\sqrt{2\pi}V(\alpha)} - \alpha, \tag{11}$$

whence we obtain

$$\overline{\tau(\alpha)} = \frac{\overline{z(\alpha)}\,\sigma}{\Delta\omega} \tag{12}$$

by taking into account the relationship between $\underline{z}$ and τ.

The probability that the duration of a peak at the level U_{th} will exceed the duration τ_0 is equal to

$$P\{\tau \geqslant \tau_0\} = P\{z \geqslant z_0\} = \frac{V(z_0 + \alpha)}{V(\alpha)}. \tag{13}$$

2. A Low-Pass Filter of the Gaussian Type

The frequency response for such a filter is written as

$$C(\omega) = e^{-\frac{\omega^2}{2\gamma^2}} \qquad (0 \leqslant \omega < \infty). \tag{14}$$

The passband of the filter at the 0.7 level is equal to $\Delta\omega = 0.84\ \gamma$.

For conditions identical to calls which were applied to the RC filter the pulse at the output of the filter is defined by the relationships

$$u_{out}(t) = \frac{C_0 x\gamma}{\sqrt{2\pi}} e^{-\frac{\gamma^2 t^2}{2}} \quad (-\infty < t < \infty). \tag{15}$$

The time τ during which the noise peak with an amplitude $\underline{x}$ exceeds the threshold voltage U_{th} is found from the equation

$$U_{th} = 0.48 C_0 \Delta\omega x e^{-\left(\frac{0.84\Delta\omega\tau}{2}\right)^2}, \tag{16}$$

By solving this equation we obtain

$$(0.42\Delta\omega\tau)^2 = \ln x - \ln U_n', \tag{17}$$

where $U_n' = \frac{U_{th}}{0.48C_0\Delta\omega}$ is the value of the threshold voltage normalized to the filter input.

Thus the quantity $(0.42\ \Delta\omega\tau)^2$ has a distribution that coincides with the distribution for ln x but with a shifted distribution center. Just as in the preceding case (8), it is convenient to express U_n' as

$$U_n' = ae^{\beta v} \tag{18}$$

and to study the normalized truncated normal distribution for $v = \frac{(0.42\Delta\omega\tau)^2}{\sigma}$:

$$p(v) = \frac{e^{-\frac{(v+\beta)^2}{2}}}{\sqrt{2\pi}V(\beta)} \quad (v \geqslant 0). \tag{19}$$

In that case the distribution for the quantity $\sqrt{v} = u$ can be written as

$$p(u) = \frac{2ue^{-\frac{(u^2+\beta)^2}{2}}}{\sqrt{2\pi}V(\beta)} \tag{20}$$

Figure 3 shows the probability distribution density curves for the duration of the peaks in the case of a low-pass Gaussian filter for a number of values of the parameter β. The maximum value for p(u) is obtained when

$$u_{max} = \sqrt{-\frac{\beta}{2} + \sqrt{\frac{\beta^2}{4} + \frac{1}{2}}}, \tag{21}$$

here the minimal value for $p(u_{max}, \beta)$ corresponds to $\beta = -\frac{1}{2}$.

The probability that the duration of a peak at the level U_{th} will exceed the specified value τ_0 is equal to (just as in the case of the RC filter)

$$P\{\tau \geqslant \tau_0\} = P\{v \geqslant v_0\} = \frac{V(v_0+\beta)}{V(\beta)}. \tag{22}$$

In order to determine the average duration of the peak at the level corresponding β we study the expression

$$\overline{u(\beta)} = \frac{2}{\sqrt{2\pi}V(\beta)} \int_0^\infty u^2 e^{-\frac{(u^2+\beta)^2}{2}}\, du. \tag{23}$$

After performing certain transformations involving the use of the substitution $u^2 = x$, formula (23) reduces to the form

$$\overline{u(\beta)} = \frac{e^{-\frac{\beta^2}{2}}}{\sqrt{2\pi}V(\beta)} \int_0^\infty x^{\frac{1}{2}} e^{-\frac{x^2}{2}-\beta x}\, dx. \tag{24}$$

The integral (24) can be expressed in terms of confluent hypergeometric functions (cf. for example, [3])

$$\int_0^\infty x^k e^{-\frac{ax^2}{2}-bx}\,dx = a^{-\frac{k+1}{2}}\,2^{\frac{k-1}{2}}\left[\Gamma\left(\frac{k+1}{2}\right){}_1F_1\left(\frac{k+1}{2},\ \frac{1}{2},\ \frac{b^2}{2a}\right)-\right.$$
$$\left.-b\sqrt{\frac{2}{a}}\,\Gamma\left(\frac{k}{2}+1\right){}_1F_1\left(\frac{k}{2}+1,\ \frac{3}{2},\ \frac{b^2}{2a}\right)\right], \tag{25}$$

where Γ is a gamma function, and ${}_1F_1$ is a confluent hypergeometric function specified in the form of curves [4]. Based on (25), we can write the following expression for $\overline{u(\beta)}$:

$$\overline{u(\beta)} = \frac{e^{-\frac{\beta^2}{2}}}{\sqrt{2\pi}\,V(\beta)}\,2^{-\frac{1}{4}}\left[\Gamma\left(\frac{3}{4}\right){}_1F_1\left(\frac{3}{4},\ \frac{1}{2},\ \frac{\beta^2}{2}\right)-\sqrt{2}\,\beta\Gamma\left(\frac{5}{4}\right){}_1F_1\left(\frac{5}{4},\ \frac{3}{2},\ \frac{\beta^2}{2}\right)\right]. \tag{26}$$

The values of $\overline{u(\beta)}$ determined from formula (26) are given in Fig. 3.

The average duration of a peak at the level corresponding to β is equal to

$$\overline{\tau(\beta)} = \frac{\overline{u(\beta)}\sqrt{\sigma}}{0.42\,\Delta\omega}. \tag{27}$$

For comparison of results Fig. 4 shows two pairs of distribution curves for the duration of the peaks at the outputs of RC and Gaussian filters with identical passbands at the level 0.7 for identical values of the threshold voltage U_{th}. It can be noted from formulas (8) and (18) that the cutoff coefficients α and β are related by the following simple expressions for identical $\Delta\omega$ and U_{th}:

$$\beta = \alpha + \frac{0.74}{\sigma}. \tag{28}$$

In order to simplify the plotting of the curves in Fig. 4 the value of σ is assumed equal to 0.74. For this case

$$\beta = \alpha + 1. \tag{29}$$

The analysis which we have performed of the distribution of the duration of the peaks for pulse noise with a logarithmically normal amplitude distribution for two types of frequency responses can easily be extended to filters with different frequency responses. Moreover, it is possible in analogous fashion to find the probability distribution density for the duration of the peaks which directly follow a detector which isolates the pulse noise envelope. Under these conditions the distribution of the durations will be determined entirely by the shape of the frequency response of the wide-band input filter.

The results obtained above make it possible to compute the actual noise immunity of remote control systems that operate over channels with pulse noise of this type when the limitations imposed on the pulse repetition frequency and the passband width of the filters are satisfied.

The problem of finding the optimal frequency response for a low-pass filter so as to insure the optimum noise immunity of a remote control system is of especial interest. This problem shall be treated in a special paper.

LITERATURE CITED

1. L. B. Venchkovskii, "Investigation of noise in a power line rated at 0.4/6 kv," Avtomatika i Telemekhanika 21, 8 (1960).
2. V. P. Pevnitskii and A. G. Frantsuz, "On the statistical distributions of the amplitudes of pulse noise signals produced by electrical devices," Elektrosviaz' 9 (1958).
3. D. Middleton, An Introduction to Statistical Communication Theory. McGraw-Hill Book Co., 1960.
4. E. Jahnke and F. Emde, Tables of Functions [Russian translation] (Fizmatigiz, 1959).

SOME CASES OF SIMILARITY OF TRANSIENT PROCESSES IN SINGLE-LOOP NONLINEAR CONTROL SYSTEMS

V. T. Dedesh

(Moscow)

Translated from Avtomatika i Telemekhanika, Vol. 22, No. 6,
pp. 801-802, June, 1961
Original article submitted December 9, 1960

Many nonlinear control systems can be represented in the form of single-loop systems that consist of two portions—a linear portion with the transfer function W(p) and a nonlinear element NE (Fig. 1).

An analysis of such control systems shows that, in the case of nonlinearities of a certain given type, similar transient processes can occur in such systems and the results of a particular experiment can be generalized.

We shall assume that the characteristics of the nonlinear element are described by the equation

$$z = f(x, \dot{x}, \delta x), \tag{1}$$

where z is the nonlinear element output quantity, x is the controlled variable, and δx is the quantity determining coordinate of the characteristic point of the NE nonlinear characteristic.

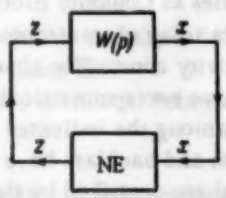


Fig. 1.

Fig. 2.

If $f(x, \dot{x}, \delta x)$ is a uniform first-power function of x, $\dot{x}$, and δx, i.e., the following equations hold:

$$f(x, \dot{x}, \delta x) = \delta x f\left(\frac{x}{\delta x}, \frac{\dot{x}}{\delta x}, 1\right) \tag{2}$$

or

$$\frac{f(x, \dot{x}, \delta x)}{\delta x} = f\left(\frac{x}{\delta x}, \frac{\dot{x}}{\delta x}, 1\right), \tag{3}$$

then, the general euqation of the system

$$W(p)\, x = f(x, \dot{x}, \delta x) \tag{4}$$

can be written in the following form:

$$W(p)\frac{x}{\delta x} = f\left(\frac{x}{\delta x}, \frac{\dot{x}}{\delta x}, 1\right). \tag{5}$$

Such a transition from the variables $\underline{x}$ and $\dot{x}$ to the variables $x / \delta x$ and $\dot{x} / \delta x$ makes it possible, in these cases, to consider a control system with a single value of the characteristic nonlinearity parameter (Fig. 2) instead of a single-loop control system with a nonlinearity having different values of the characteristic parameter δx.

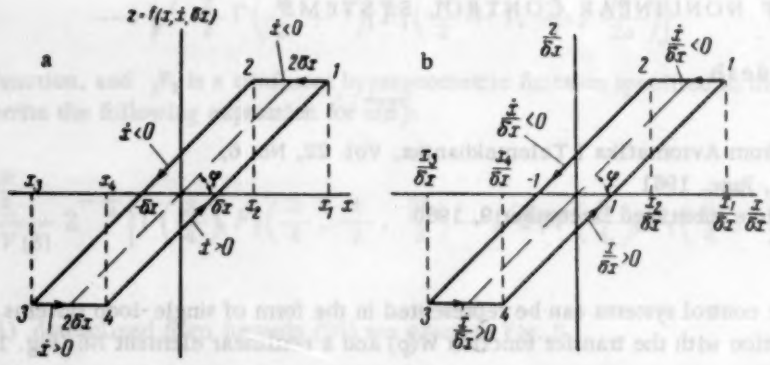


Fig. 3.

Consequently, in single-loop control systems consisting of a linear portion and a nonlinear element that is given by $f(x, \dot{x}, \delta x)$, the separation of a group of similar transient processes is possible if the nonlinearity is described by a uniform first-power function of $\underline{x}$, $\dot{x}$, δx. Here, the similarity criteria are the ratios of the controlled variable $\underline{x}$ and of its derivative $\dot{x}$ to the coordinates δx of the nonlinearity characteristic point.

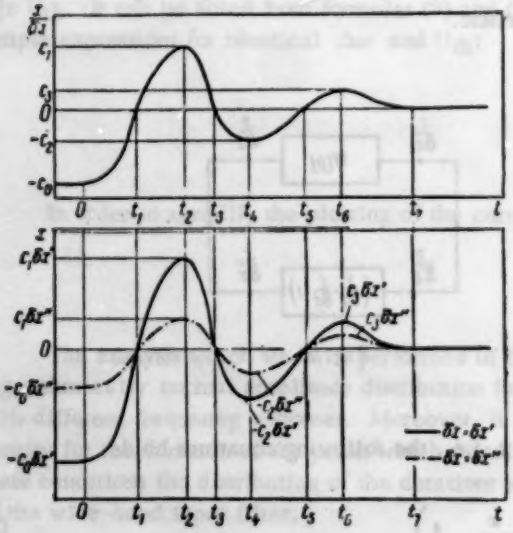


Fig. 4.

In particular, uniform functions are used to describe such nonlinearities as Coulomb friction, backlash, nonlinearities with a relay characteristic, and also nonlinearities with an insensitivity zone. The characteristics of the above nonlinearities have two symmetrical characteristic points (Fig. 3). From among the indicated nonlinearities, only Coulomb friction and backlash have characteristics that depend on $\dot{x}$ and are described by the equation

$$z = f(x, \operatorname{sign} \dot{x}, \delta x). \tag{6}$$

Nonlinearities with an insensitivity zone and with a relay characteristic are described by the equation

$$z = f(x, \delta x). \tag{7}$$

Under specific conditions, uniform equations can also be used for describing nonlinearities that have a relay characteristic with an insensitivity zone as well as a characteristic with limitation or with limitation and an insensitivity zone. The sufficient uniformity condition for these dependences is the simultaneous proportional variation of all the characteristic points of a nonlinearity, which secures the invariability of the nonlinearity in switching over to the $x / \delta x$ variable.

In the cases where $x/\delta x$ and $\dot{x}/\delta x$ are used as similarity criteria, the transition to new variables, which are formed by dividing the controlled variable $\underline{x}$ and its derivative $\dot{x}$ by the value δx of this quantity at the nonlinearity characteristic point, makes it possible to arrive at the following general conclusions that are valid for the conditions considered.

1. Within a group of similar transient processes and during equal time intervals, the values of the controlled variables x' and x" are expressed by the $\delta x' / \delta x''$ ratio, while the time interval remains unchanged within the group of similar transient processes (Fig. 4).

A consequence of this conclusion is the fact that the self-oscillation amplitude is proportional to the δx value and the fact that the oscillation period is independent of the δx value.

2. The character of the sequence of limiting cycles and the character of the trajectory of the point in the space between two limiting cycles of the phase diagram are independent of the δx value. Hence, it follows that the stability boundaries of the control system in the domain of the linear portion parameters are independent on the δx value and that they are determined only by the nonlinearity type.

3. In analyzing transient processes in similar single-loop nonlinear control systems of the type considered here by means of phase diagrams, it is advisable to use, as the coordinates, the ratios of the controlled variable and of its derivative to the coordinate δx of the nonlinearity characteristic point.

The author extends his thanks to Prof. V. S. Vedrov for his advice, which made it possible to present this communication in the above form.

ON THE UTILIZATION OF DIGITAL COMPUTERS FOR DIFFERENTIATING AND SMOOTHING OUT SEQUENCES CONTAINING RANDOM DISTURBANCES

A. N. Pokrovskii

(Leningrad)

Translated from Avtomatika i Telemekhanika, Vol. 22, No. 6,
pp. 803-806, June, 1961
Original article submitted February 29, 1960

The present article emphasizes the fact that a large number of arithmetical operations must be performed in using digital computers (DC) for solving the problem of optimum smoothing-out and for effecting linear transformations of sequences of signals containing random disturbances.

It is demonstrated that, in a number of cases, the necessary number of arithmetical operations can be considerably reduced by using the recursion formula.

In using digital computers in automatic control systems, it often becomes necessary to perform smoothing-out, differentiation, or any other linear operation on the sequence of values of a signal containing a disturbance. The desired transformation can be brought about in the following manner by using sliding summation.

We shall assume that the signal x(t) values are read at equal time intervals T_0 : $x_k = x(kT_0)$, where k = ..., -1, 0, 1, 2, 3, ... By using a linear combination of the last N + 1 readings

$$y_k = \sum_{n=0}^{N} w_n x_{k-n} \tag{1}$$

and for a suitable choice of weighting coefficients w_n, we can approximate any linear transformation by the useful component of the signal.

By using the criterion for the minimum of the root-mean-square error, we arrive at a system of linear equations for determining the optimum values of the weighting coefficients w_n:

$$\sum_{n=0}^{N} R(jT_0 - nT_0)\, w_n = F(j) \qquad (j = 0, 1, 2, \ldots, N), \tag{2}$$

and also at some additional conditions [1-3].

In Eqs (2), F(j) is the linear combination of certain functions, and $R(\tau)$ is the autocorrelation function of the steady-state random signal. We shall assume that the signal random components are such that the spectral function corresponding to the autocorrelation function $R(\tau)$ is a rational-fraction function:

$$f(\zeta) = \sum_{n=-\infty}^{\infty} R(nT_0)\,\zeta^{-n} = \frac{A(\zeta)}{B(\zeta)},$$
$$A(\zeta) = \sum_{j=-m}^{m} a_j \zeta^j, \qquad B(\zeta) = \sum_{j=-l}^{l} b_j \zeta^j. \tag{3}$$

In this case, as was shown in [5], the sequence of weighting coefficients w_n can be expressed by

$$w_n = \varphi(n) + \sum_{s=1}^{2m} B_s \Phi_s(n) + \sum_{s=0}^{l-m-1} [C_s \delta_{n,s} + D_s \delta_{n,N-s}], \tag{4}$$

where $\Phi_s(n)$ is the complete system of solutions of the homogeneous difference equation

$$A(z)\,\Phi(n) = \sum_{j=-m}^{m} a_j \Phi(n+j) = 0, \tag{5}$$

and

$$\delta_{n,s} = \begin{cases} 1 & \text{for } n = s, \\ 0 & \text{for } n \neq s. \end{cases}$$

In Eq. (5) and below, $\underline{z}$ designates the unit shift operator:

$$\underline{z} x_n = x_{n+1}.$$

The $\varphi(n)$ function in Eq. (4) is a linear combination of certain functions. If the useful component of the signal is approximated by a polynomial with the power r, $\varphi(n)$ is a polynomial of the same power:

$$\varphi(n) = \gamma_0 + \gamma_1 n + \ldots + \gamma_r n^r. \tag{6}$$

If the optimum values of the weighting coefficients w_n are known, the sequence y_k ($k = \ldots, 1, 2, 3, \ldots$) can be calculated by using Eq. (1). However, for large N values, a large number of arithmetical operations (N + 1 multiplications and N additions) must be performed at each calculation step. This would impose severe requirements on the DC speed.

As will be shown below, the number of arithmetical operations necessary for calculating y_k can be reduced at the expense of a certain increase in the utilized volume of the operative memory (OM) in DC.

Assume that $\varphi(n)$ satisfies the difference equation

$$E(z)\,\varphi(n) = \sum_{k=0}^{p} e_k \varphi(n+k) = 0 \qquad (e_0 \neq 0,\ e_p \neq 0). \tag{7}$$

Then, if

$$G(z) = A(z)\,E(z) = \sum_{s=-m}^{p+m} g_s z^s, \tag{8}$$

where

$$g_s = \sum_{k=0}^{p} a_{s-k} e_k \tag{9}$$

($a_k = 0$ for $k > m$ and for $k < -m$),

$$G(z)\,\varphi(n) = 0, \quad G(z)\,\Phi_s(n) = 0. \tag{10}$$

We shall apply the $z^{-p-m} G(z)$ operator to both sides of Eq. (1); as a result of this, we shall have:

$$z^{-p-m} G(z)\, y_k = \sum_{n=0}^{N} w_n G(z)\, z^{-p-m} x_{k-n}$$

or

$$\text{By substituting the summation index} \quad \sum_{s=-m}^{p+m} g_s y_{k+s-p-m} = \sum_{n=0}^{N} \sum_{s=-m}^{p+m} g_s w_n x_{k-n-p-m+s}. \tag{11}$$

we find

$$n = s + \alpha, \quad -m-p \leqslant \alpha \leqslant N+m,$$

$$\sum_{s=-m}^{p+m} g_s y_{k-p-m+s} = \sum_{\alpha=-m-p}^{N+m} x_{k-p-m-\alpha} \sum_{s=-p}^{p+m} g_s w_{s+\alpha}. \tag{12}$$

Let us denote

$$d_\alpha = \sum_{s=-m}^{p+m} g_s w_{s+\alpha} \tag{13}$$

and divide the region where α changes into five intervals and then calculate d_α for each of these intervals:

$$\begin{aligned}
&1)\ -m-p \leqslant \alpha \leqslant m-1; \quad d_\alpha = \sum_{s=-\alpha}^{p+m} g_s w_{s+\alpha},\\
&2)\ m \leqslant \alpha \leqslant l-1; \quad d_\alpha = \sum_{s=0}^{l-m-1} C_s g_{s-\alpha},\\
&3)\ l \leqslant \alpha \leqslant N-p-l; \quad d_\alpha = 0,\\
&4)\ N-l-p+1 \leqslant \alpha \leqslant N-p-m; \quad d_\alpha = \sum_{s=0}^{l-m-1} D_s g_{N-\alpha-s},\\
&5)\ N-p-m+1 \leqslant \alpha \leqslant N+m; \quad d_\alpha = \sum_{s=-m}^{N-\alpha} g_s w_{s+\alpha}.
\end{aligned} \tag{14}$$

The calculation of d_α in the first and the fifth interval does not require an explanation. In the third interval, $d_\alpha = 0$ because of condition (10).

In the second and the fourth interval, d_α is expressed in terms of the coefficients in front of the δ-functions in Eq. (4). In calculating the sums, it must be borne in mind that $g_s = 0$ for $s < -m$ and $s > p + m$. Finally, we obtain:

$$\sum_{s=-m}^{p+m} g_s y_{k-p-m-s} = \sum_{\alpha=-m-p}^{l-1} d_\alpha x_{k-p-m-\alpha} + \sum_{\alpha=N-l-p+1}^{N+m} d_\alpha x_{k-p-m-\alpha}. \quad (15)$$

If we divide all terms by g_{m+p} ($g_{m+p} = a_m e_p \neq 0$) and transfer all terms with the exception of y_k to the right-hand side, we find:

$$y_k = -\sum_{s=-m}^{p+m-1} \frac{g_s}{g_{m+p}} y_{k-p-m+s} + \sum_{\alpha=-m-p}^{l-1} \frac{d_\alpha}{g_{m+p}} x_{k-p-m-\alpha} + \sum_{\alpha=N-l-p+1}^{N+m} \frac{d_\alpha}{g_{m+p}} x_{k-p-m-\alpha} \quad (16)$$

Equation (16) provides the possibility of finding y_k by using a recurrent process suitable for programming. If Eq. (16) is used, the number of calculations does not depend on N, and, for large N values, it can be considerably smaller than the number of calculations where Eq. (1) is used.

In the particular case where $m = l = 0$ and $E(z) = (1 - z)^p$, Eq. (16) is transformed into the equation which was obtained in a different manner in [4].

Example. We shall assume that the autocorrelation function can be represented in the form

$$R(\tau) = e^{-\alpha|\tau|}[\cos\beta\tau + c\sin\beta|\tau|], \quad (17)$$

while the "regular" signal can be expressed by a second-power polynomial. In this case, the spectral function is given by

$$f(z) = \theta_0 \frac{A(z)}{B(z)}, \quad (18)$$

where

$$\theta_0 = \operatorname{sh} 2\alpha T_0 (\bar{a}_1 - \bar{a}_2 c), \quad A(z) = 2\operatorname{ch}\rho - \left(z + \frac{1}{z}\right),$$

$$B(z) = \left[2\operatorname{ch}\gamma T_0 - \left(z + \frac{1}{z}\right)\right]\left[2\operatorname{ch}\bar{\gamma} T_0 - \left(z + \frac{1}{z}\right)\right],$$

$$\operatorname{ch}\rho = 1 + \frac{(1 - \bar{a}_1)(1 + c_2\bar{a}_2)}{\bar{a}_1 - c_2\bar{a}_2}. \quad (19)$$

$$\gamma = \alpha - i\beta, \quad \bar{\gamma} = \alpha - i\beta, \quad \bar{a}_1 = \frac{\cos\beta T_0}{\operatorname{ch}\alpha T_0}, \quad \bar{a}_2 = \frac{\sin\beta T_0}{\operatorname{sh}\alpha T_0}.$$

We further find:

$$m = 1, \ l = 2, \ \Phi_1(n) = e^{n\rho}, \ \Phi_2(n) = e^{-n\rho},$$
$$a_{-1} = -1, \ a_0 = 2\operatorname{ch}\rho, \ a_1 = -1,$$
$$\varphi(n) = \varphi_0 + \varphi_1 n + \varphi_2 n^2,$$
$$E(z) = (1 - z)^3,$$

$$p = 3, \ e_0 = 1, \ e_1 = -3, \ e_2 = 3, \ e_3 = -1,$$
$$g_{-1} = -1, \ g_0 = 3 + 2\operatorname{ch}\rho, \ g_1 = -4 - 6\operatorname{ch}\rho,$$
$$g_2 = 4 + 6\operatorname{ch}\rho, \ g_3 = -3 - 6\operatorname{ch}\rho, \ g_4 = 1,$$
$$d_{-4} = g_4 w_0,$$
$$d_{-3} = g_3 w_0 + g_4 w_1,$$
$$d_{-2} = g_2 w_0 + g_3 w_1 + g_4 w_2,$$
$$d_{-1} = g_1 w_0 + g_2 w_1 + g_3 w_2 + g_4 w_3,$$
$$d_0 = g_0 w_0 + g_1 w_1 + g_2 w_2 + g_3 w_3 + g_4 w_4,$$
$$d_1 = g_{-1} C_0,$$
$$d_{N-4} = g_4 D_0,$$
$$d_{N-3} = g_0 w_{N-3} + g_1 w_{N-2} + g_2 w_{N-1} + g_3 w_N + g_{-1} w_{N-4},$$
$$d_{N-2} = g_0 w_{N-2} + g_1 w_{N-2} + g_2 w_N + g_{-1} w_{N-3},$$
$$d_{N-1} = g_0 w_{N-1} + g_1 w_N + g_{-1} w_{N-2},$$
$$d_N = g_{-1} w_{N-1} + g_0 w_N,$$
$$d_{N+1} = g_{-1} w_N.$$

Equation (16) assumes the following form:

$$y_k = -[g_{-1}y_{k-5} + g_0y_{k-4} + g_1y_{k-3} + g_2y_{k-2} + g_3y_{k-1}] + d_{-4}x_k + d_{-3}x_{k-1} + d_{-2}x_{k-2} + d_{-1}x_{k-3} + d_0x_{k-4} + d_1x_{k-5} + d_{N-4}x_{k-N} + d_{N-3}x_{k-N-1} + d_{N-2}x_{k-N-2} + d_{N-1}x_{k-N-3} + d_Nx_{k-N-4} + d_{N+1}x_{k-N-5}. \quad (20)$$

The latter equation is suitable for programming. In calculations, N + 6 OM cells are required for storing the $x_k \div x_{k-N-5}$ quantities and six cells are required for the $y_k \div y_{k-5}$ quantities; in all, N + 12 OM cells are required.

Moreover, 17 coefficients must be stored in the memory: $g_{-1} \div g_3$; $d_{-4} \div d_1$; $d_{N-4} \div d_{N+1}$. For the performance of each calculation cycle, 17 multiplications and 16 additions are necessary, while in calculations where Eq. (1) is used, for instance, for N + 1 = 100, 100 multiplications and 99 additions must be performed.

LITERATURE CITED

1. K. Jonson, Optimum, Linear, Discrete Filtering of Signals, Containing a Nonrandom Component. Trans. IRE Vol. 1, T−2, No. 2, June, 1956.
2. V. P. Perov, Statistical Synthesis of Pulse Systems [in Russian] (Izd. Sov. Radio, 1959).
3. Ya. Z. Tsypkin, Theory of Pulse Systems [in Russian] (Fizmatgiz, 1958).
4. M. Blum, Optimum, Discrete, Mean Square Error Filtering Using Recursion Formula..Trans. IRE Vol. 1, T−3, No. 3, September, 1957.
5. A. N. Pokrovskii, "On the calculation of discrete filters." Theory and Application of Discrete Automatic Systems [in Russian] (Izd. AN SSSR, 1960).

ON TRANSIENT RESPONSE CONDITIONS IN A MAGNETIC AMPLIFIER WITH FEEDBACK AND AN INDUCTIVE LOAD THAT IS CONNECTED THROUGH A RECTIFIER

V. A. Sokolov

(Rostov-on-Don)

Translated from Avtomatika i Telemekhanika, Vol. 22, No. 6, pp. 807-810, June, 1961
Original article submitted March 8, 1961

The present article is concerned with a method for investigating transient response conditions in a magnetic amplifier with internal and external feedbacks, which is encompassed by flexible feedback and which has an inductive load that is connected through a rectifier.

The problem of transient response conditions in magnetic amplifiers without feedback for an inductive load that is connected through a rectifier bridge has been treated in sufficient detail in [1, 2]. These investigations were concerned with amplifiers using cores and rectifiers with idealized [1] as well as nonidealized [2] characteristics.

Amplifiers with a more complicated circuit are very often used in automatic control systems. Therefore, it would be of definite interest to determine to what extent can the method proposed in [2] for investigating transient response conditions be extended to the more general case where a magnetic amplifier has several control windings, internal or external positive feedback, flexible feedback, and an inductive load that is connected through a rectifier bridge. We shall limit our consideration to the weak signal regime in the magnetic amplifier, i.e., to the

regime of such signals for which the currents and voltages acting in the amplifier's working windings do not exceed the limits of the rectilinear portions of the magnetization characteristics under transient as well as steady-state conditions, in connection with which the amplifier parameters can be considered as constant.

1. Transient Response Conditions in a Magnetic Amplifier with Positive External and Flexible Feedbacks and with an Inductive Load that is Connected through a Rectifier

Consider a magnetic amplifier which has two control windings W_{c1} and W_{c2}, to which control signals with different polarities are supplied, a flexible feedback winding W_{c3}, and a positive feedback winding W_{fb}. The amplifier circuit is shown in figure 1. Here, r_g and C_g are the resistance and the capacitance in the flexible feedback circuit, respectively, and r_1 and L_1 are the load resistance and inductance, respectively. The excitation winding of a dc motor represents the load.

By using the assumptions made in [2, 3], we shall compose the equivalent amplifier circuit, which is shown in figure 2. In this circuit, in correspondence with the designations and terminology used in [2, 3], r_{c1}, r_{c2}, r_{c3}, and r_{fb} are the resistance of the W_{c1}, W_{c2}, W_{c3}, and W_{fb} windings; r_a and r_b are the forward and the reverse resistances of the rectifier arm; $R_i' = \dfrac{(R_i + r_a)\,0.5 r_b}{R_i + r_a + 0.5 r_b}$; R_i is the amplifier's internal resistance; $S' = \dfrac{R_i}{R_i + R_a} S$, where S is the amplifier response slope; ε and ρ are coefficients which take into account the effect of commutation in rectification; $L_1' = L_1/\rho$; $\beta = W_{fb}/W_c$; $\beta_c = W_{c3}/W_c$; $W_{c1} = W_{c2} = W_c$; Δi_{c1}, Δi_{c2}, Δi_{c3}, Δi, Δi_1 are the current increments in the corresponding amplifier circuits with respect to the initial current values.

Let us consider the amplifier transient response equations for the case where the control voltages change in steps. Considering that, in core magnetization, the increments of the voltage at the rectifier input terminals and of the emf of the ac windings are equal and have opposite signs, the expression for the increment in the magnetic induction constant component for the equivalent circuit under consideration can be written as

$$\Delta B_0 = \frac{\mu_{c0} W_c}{l}(\Delta i_{c2} - \Delta i_{c1} + \beta \Delta i_1 - \beta_c \Delta i_{c3}) + \\ + \frac{\nu}{\omega} \frac{r_1' \Delta i_1 + W_{fb} Q \dfrac{d\Delta B_0}{dt} + L_1' \dfrac{d\Delta i_1}{dt}}{W_\sim Q}, \qquad (1)$$

where Q is the cross-sectional area of the two cores, l is the average length of the magnetic lines of force, $\mu_c = \partial B_0/\partial H_0$ for $B_\sim = \text{const}$, $\nu = -\partial B_0/\partial B_\sim$ for $H_0 = \text{const}$, and ω is the alternating current angular frequency.

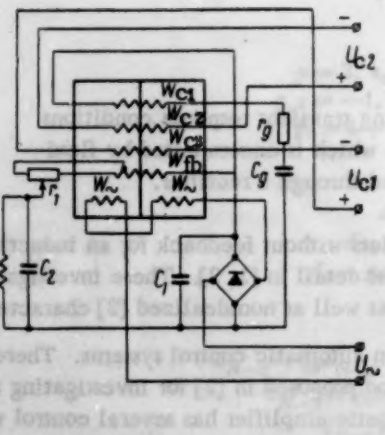


Fig. 1.

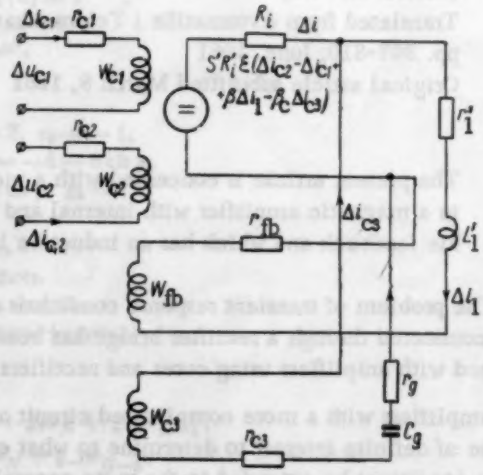


Fig. 2.

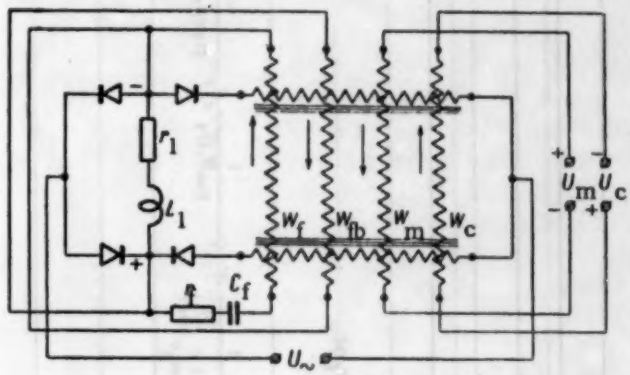


Fig. 3.

On the basis of the equivalent circuit for the amplifier output circuit, we can write the following equation:

$$S'R_i'\xi(\Delta i_{c2}-\Delta i_{c1}+\beta\Delta i_1-\beta_c\Delta i_{c3})=R_i'(\Delta i_1+\Delta i_{c3})+r_1'\Delta i_1+L_1'\frac{d\Delta i_1}{dt}+W_{fb}Q\frac{d\Delta B_\delta}{dt}. \quad (2)$$

The equations for the control winding circuits will have the following form:

$$\Delta u_{c1}=r_{c1}\Delta i_{c1}-W_cQ\frac{d\Delta B_\delta}{dt}, \quad (3)$$

$$\Delta u_{c2}=r_{c2}\Delta i_{c2}+W_cQ\frac{d\Delta B_\delta}{dt} \quad (4)$$

and for the flexible feedback circuit, we have:

$$S'R_i'\xi(\Delta i_{c2}-\Delta i_{c1}+\beta\Delta i_1-\beta_c\Delta i_{c3})-R_i'(\Delta i_1+\Delta i_{c3})=$$
$$=(r_g+r_{c3})\Delta i_{c3}+\frac{1}{C}\int\Delta i_{c3}dt+W_{c3}Q\frac{d\Delta B_\delta}{dt}. \quad (5)$$

Equations (1) - (5) completely describe the transient response conditions in the magnetic amplifier under consideration. After some transformations of Eqs. (1) - (5), we obtain a single equation for the transient response conditions in the magnetic amplifier under consideration:

$$a_3\frac{d^3\Delta i_1}{dt^3}+a_2\frac{d^2\Delta i_1}{dt^2}+a_1\frac{d\Delta i_1}{dt}+a_0\Delta i_1=$$
$$=b_2\frac{d^2\Delta u_{c2}}{dt^2}+b_1\frac{d\Delta u_{c2}}{dt}+b_0\Delta u_{c2}-\left(c_2\frac{d^2\Delta u_{c1}}{dt^2}+c_1\frac{d\Delta u_{c1}}{dt}+c_0\Delta u_{c1}\right). \quad (6)$$

2. Transient Response Conditions in a Magnetic Amplifier with Internal Feedback and an Inductive Load Connected to a Rectifier

At the present time, magnetic amplifiers with internal feedback are extensively used in automatic control systems. Therefore, it is of definite interest to determine to what an extent can the method for investigating transient responses that has been proposed in [2] be extended to the more general case where a magnetic amplifier with nonidealized characteristics has internal feedback, external voltage feedback, flexible feedback, and an inductive load that is connected through a rectifier bridge.

Coefficient	Expression for the coefficient
A_3	$a_1b_4 - a_4b_1$
A_2	$a_1b_3 + a_2b_4 - a_4b_2 - a_3b_1$
A_1	$a_2b_3 + a_3b_4 - a_4b_3 - a_3b_2$
A_0	$a_3b_3 - a_3b_3$
B_2	$a_4c_1 - c_3a_1$
B_1	$a_4c_2 + a_3c_1 - a_2c_3$
B_0	$a_3c_2 - a_3c_3$
a_1	$W_{\sim}QC_g - \frac{W_cQ}{\beta_2 r_c} r_gC_g + \frac{\beta_3}{\beta_2}\frac{S'\xi}{1+\frac{\beta_3}{\beta_2}}\frac{\nu l}{\omega\mu_{co}W_c} r_gC_g + \frac{\beta_3}{\beta_2}\frac{r_gC_g}{1+\frac{\beta_3}{\beta_2}}\frac{W_cQ}{\beta_2 r_c} - \frac{\beta_3}{\beta_2}\frac{r_gC_g}{1+\frac{\beta_3}{\beta_2}}\frac{\nu l}{\beta_2\omega\mu_{co}W_c} + \frac{\beta_3}{\beta_2}\frac{W_{\sim}Qr_gC_g}{R_i'\left(1+\frac{\beta_3}{\beta_2}\right)} + \frac{\nu l r_gC_g}{\beta_2\omega\mu_{co}W_c} - W_fQC_g$
a_2	$-\frac{W_cQ}{\beta_2 r_c} - \frac{\beta_3}{\beta_2}\frac{S'\xi}{1+\frac{\beta_3}{\beta_2}}\frac{lr_gC_g}{\mu_{co}W_c} + \frac{\beta_3}{\beta_2}\frac{S'\xi}{1+\frac{\beta_3}{\beta_2}}\frac{\nu l}{\omega\mu_{co}W_c} + \frac{\beta_3}{\beta_2}\frac{1}{1+\frac{\beta_3}{\beta_2}}\frac{W_cQ}{\beta_2 r_c} + \frac{\beta_3}{\beta_2}\frac{r_gC_g}{1+\frac{\beta_3}{\beta_2}}\frac{l}{\mu_{co}W_c\beta_2} - \frac{\beta_3}{\beta_2}\frac{1}{1+\frac{\beta_3}{\beta_2}}\frac{\nu l}{\beta_2\omega\mu_{co}W_c} + \frac{\beta_3}{\beta_2}\frac{1}{R_i'\left(1+\frac{\beta_3}{\beta_2}\right)} W_{\sim}Q - \frac{l}{\mu_{co}W_c\beta_2} r_gC_g + \frac{\nu l}{\beta_2\omega\mu_{co}W_c}$
a_3	$-\frac{\beta_3}{\beta_2}\frac{S'\xi}{1+\frac{\beta_3}{\beta_2}}\frac{l}{\mu_{co}W_c} + \frac{\beta_3}{\beta_2}\frac{1}{1+\frac{\beta_3}{\beta_2}}\frac{l}{\mu_{co}W_c\beta_2} - \frac{l}{\mu_{co}W_c\beta_2}$
a_4	$W_{\sim}Q + \frac{r_{fb}S'\xi}{1+\frac{\beta_3}{\beta_2}}\frac{\nu l}{\omega\mu_{co}W_c} + \frac{r_{fb}}{1+\frac{\beta_3}{\beta_2}}\frac{W_cQ}{\beta_2 r_c} - \frac{r_{fb}}{R_i'\left(1+\frac{\beta_3}{\beta_2}\right)}\frac{\nu l}{\beta_2\omega\mu_{co}W_c} + \frac{r_{fb}}{R_i'\left(1+\frac{\beta_3}{\beta_2}\right)} W_{\sim}Q - W_{fb}Q$
a_5	$-\frac{r_{fb}S'\xi}{1+\frac{\beta_3}{\beta_2}}\cdot\frac{l}{\mu_{co}W_c} + \frac{r_{fb}}{1+\frac{\beta_3}{\beta_2}}\frac{l}{\mu_{co}W_c\beta_2}$

b_1	$-L_1'C_g-\frac{\beta_3}{\beta_2}\frac{S'\xi}{1+\frac{\beta_3}{\beta_2}}\frac{\nu l}{\omega W_{\sim}Q\mu_{co}W_c}L_1'r_gC_g+\frac{\beta_3}{\beta_2}\frac{L_1'r_gC_g}{1+\frac{\beta_3}{\beta_2}}\frac{\nu l}{\beta_2\omega W_{\sim}Q\mu_{co}W_c}-\frac{\nu l}{\beta_2\omega W_{\sim}Q\mu_{co}W_c}L_1'r_gC_g-\frac{\beta_3}{\beta_2}\frac{1}{R_i'\left(1+\frac{\beta_3}{\beta_2}\right)}L_1'r_gC_g$
b_2	$-r_1'C_g-\frac{\beta_1}{\beta_2}r_gC_g-\frac{\beta_3}{\beta_2}\frac{S'\xi r_g'r_gC_g}{1+\frac{\beta_3}{\beta_2}}\frac{\nu l}{\omega W_{\sim}Q\mu_{co}W_c}-\frac{\beta_3}{\beta_2}\frac{S'\xi L_1'}{1+\frac{\beta_3}{\beta_2}}\frac{\nu l}{\omega W_{\sim}Q\mu_{co}W_c}-\frac{\beta_3}{\beta_2}\frac{r_gC_g}{1+\frac{\beta_3}{\beta_2}}+\frac{\beta_3}{\beta_2}\frac{r_gC_g}{1+\frac{\beta_3}{\beta_2}}\frac{\beta_1}{\beta_2}+\frac{\beta_3}{\beta_2}\frac{r_1'r_gC_g}{1+\frac{\beta_3}{\beta_2}}\frac{\nu l}{\beta_2\omega W_{\sim}Q\mu_{co}W_c}+\frac{\beta_3}{\beta_2}\frac{L_1'}{1+\frac{\beta_3}{\beta_2}}\frac{\nu l}{\beta_2\omega W_{\sim}Q\mu_{co}W_c}-$ $-\frac{\beta_3}{\beta_2}\frac{r_1'r_gC_g}{R_i'\left(1+\frac{\beta_3}{\beta_2}\right)}-\frac{\beta_3}{\beta_2}\cdot\frac{L_1'}{R_i'\left(1+\frac{\beta_3}{\beta_2}\right)}-\frac{\nu l r_1'r_gC_g}{\beta_2\omega W_{\sim}Q\mu_{co}W_c}-\frac{\nu l L_1'}{\beta_2\omega W_{\sim}Q\mu_{co}W_c}$
b_3	$-\frac{\beta_1}{\beta_2}-\frac{\beta_3}{\beta_2}\frac{S'\xi r_1'}{1+\frac{\beta_3}{\beta_2}}\frac{\nu l}{\omega W_{\sim}Q\mu_{co}W_c}-\frac{\beta_3}{\beta_2}\frac{1}{1+\frac{\beta_3}{\beta_2}}+\frac{\beta_3}{\beta_2}\frac{1}{1+\frac{\beta_3}{\beta_2}}\frac{\beta_1}{\beta_2}+\frac{\beta_3}{\beta_2}\frac{1}{1+\frac{\beta_3}{\beta_2}}\frac{\nu l r_1'}{\beta_2\omega W_{\sim}Q\mu_{co}W_c}-\frac{\beta_3}{\beta_2}\frac{r_1'}{R_i'\left(1+\frac{\beta_3}{\beta_2}\right)}-\frac{\nu l r_1'}{\beta_2\omega W_{\sim}Q\mu_{co}W_c}$
b_4	$-L'-\frac{r_{fb}S'\xi}{1+\frac{\beta_3}{\beta_2}}\frac{\nu l L_1'}{\omega W_{\sim}Q\mu_{co}W_c}-\frac{r_{fb}}{1+\frac{\beta_3}{\beta_2}}\frac{\nu l L_1'}{\beta_2\omega W_{\sim}Q\mu_{co}W_c}-\frac{r_{fb}L_1'}{R_i'\left(1+\frac{\beta_3}{\beta_2}\right)}$
b_5	$-r_1'-\frac{r_{fb}}{1+\frac{\beta_3}{\beta_2}}\frac{S'\xi\nu l r_1'}{\omega W_{\sim}Q\mu_{co}W_c}-\frac{r_{fb}}{1+\frac{\beta_3}{\beta_2}}\left(1-\frac{\beta_1}{\beta_2}\right)\cdot\frac{r_{fb}}{1+\frac{\beta_3}{\beta_2}}\frac{\nu l r_1'}{\beta_2\omega W_{\sim}Q\mu_{co}W_c}-\frac{r_{fb}r_1'}{R_i'\left(1+\frac{\beta_3}{\beta_2}\right)}$
c_1	$-\frac{r_gC_g}{\beta_3r_C}-\frac{\beta_3}{\beta_2}\frac{1}{1+\frac{\beta_3}{\beta_2}}\frac{r_gC_g}{\beta_3r_C}$
c_2	$-\frac{1}{\beta_3r_C}+\frac{\beta_3}{\beta_2}\frac{1}{1+\frac{\beta_3}{\beta_2}}\frac{1}{\beta_3r_C}$
c_3	$\frac{r_{fb}}{1+\frac{\beta_3}{\beta_2}}\frac{1}{\beta_3r_C}$

The circuit of such an amplifier is shown in figure 3.

The magnetic amplifier has voltage feedback, which is realized by means of the W_{fb} winding. The amplifier is encompassed by flexible feedback, which comprises resistor r_f, capacitor C_f, and the winding W_f of the magnetic amplifier. Moreover, the magnetic amplifier also has a magnetization winding, W_m, which is connected to an additional supply source. The excitation winding of a dc motor, which has resistance r_1 and inductance L_1, serves as the amplifier load.

By using the assumptions made in [2, 3], we arrive at the amplifier equivalent circuit that is shown in figure 4. Here, the effect of the magnetization winding W_m on the transient response conditions in the amplifier is not taken into account. In the circuit of figure 4,

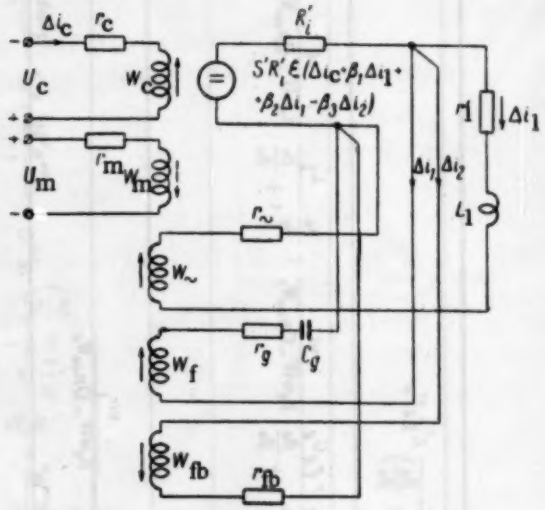


Fig. 4.

$$r_1' = r_1 + r_a,\ r_g = r_f + r_{W_f},\ C_g = C_f,\ L_1' = L_1/p,\ \beta_1 = W_\sim/W_g,\ \beta_2 = W_f/W_c,\ \beta_3 = W_{fb}/W_c.$$

The expression for the increment in the magnetic induction constant component for the equivalent circuit under consideration can be written as

$$\Delta B_0 = \frac{\mu_{c0} W_c}{l}(\Delta i_c + \beta_1 \Delta i_1 + \beta_2 \Delta i_1 - \beta_3 \Delta i_2) + \frac{\nu}{\omega W_\sim Q}\left(r_1' \Delta i_1 + L_1' \frac{d\Delta i_1}{dt} + W_\sim Q \frac{d\Delta B_0}{dt}\right). \tag{7}$$

We shall now write the equations for the amplifier circuits:

output circuit equation:

$$S' R_i' \xi (\Delta i_c + \beta_1 \Delta i_1 + \beta_2 \Delta i_1 - \beta_3 \Delta i_2) = R_i'(\Delta i_1 + \Delta i_1 + \Delta i_2) + r_1' \Delta i_1 + L_1' \frac{d\Delta i_1}{dt} + W_\sim Q \frac{d\Delta B_0}{dt}; \tag{8}$$

control circuit equation:

$$\Delta u_c = r_c \Delta i_c + W_c Q \frac{d\Delta B_0}{dt}; \tag{9}$$

equation for the flexible feedback circuit:

$$S' R_i' \xi (\Delta i_1 + \beta_1 \Delta i_1 + \beta_2 \Delta i_1 - \beta_3 \Delta i_2) - R_i'(\Delta i_1 + \Delta i_1 + \Delta i_2) = r_g \Delta i_1 + \frac{1}{C_g}\int \Delta i_1 dt + W_f Q \frac{d\Delta B_0}{dt}; \tag{10}$$

equation for the voltage feedback circuit:

$$S' R_i' \xi (\Delta i_c + \beta_1 \Delta i_1 + \beta_2 \Delta i_1 - \beta_3 \Delta i_2) - R_i'(\Delta i_1 + \Delta i_1 + \Delta i_2) = r_{fb} \Delta i_2 + W_{fb} Q \frac{d\Delta B_0}{dt}. \tag{11}$$

The system of equations (7) - (11) completely describes the transient response conditions prevailing in the amplifier under consideration. After performing a number of transformations with these equations, we obtain the transient response equation:

$$A_3 \frac{d^3 \Delta i_1}{dt^3} + A_2 \frac{d^2 \Delta i_1}{dt^2} + A_1 \frac{d\Delta i_1}{dt} + A_0 \Delta i_1 = B_2 \frac{d^2 \Delta u_c}{dt^2} + B_1 \frac{d\Delta u_c}{dt} + B_0 \Delta u_c. \tag{12}$$

The expressions for the coefficients in Eq. (12) are given in the table.

We used Eqs. (6) and (12) for calculating the transient responses in actual amplifier specimens, after which we obtained oscillograms of the transient responses in the amplifiers under investigation for a check of the calculation data.

A comparison between the theoretical and experimental data confirmed the correctness of our calculations and conclusions.

SUMMARY

1. The performed calculations and experiments confirmed the possibility of using the method for investigating transient response conditions that has been proposed in [2] for the majority of magnetic amplifier circuits with internal feedback. Regardless of the great amount of calculation involved in determining the coefficients of the equations, this method is sufficiently simple and clear.

2. The obtained equations for the calculation of transient response conditions in fairly complicated amplifiers with feedback make it possible to estimate the effect of the parameters of amplifiers and rectifier bridges on the duration of transient response conditions in an inductive load for weak signals.

LITERATURE CITED

1. N. P. Vasil'eva, "Operation of magnetic amplifiers with inductive-resistive load that is connected through a rectifier bridge," Avtomatika i Telemakhnika Vol. 15, No. 6, 1954.
2. L. V. Safris, "On the problem of transient response conditions in magnetic amplifiers with inductive loads that are connected through a rectifier," Avtomatika i Telemekhanika Vol. 19, No. 3, 1958.
3. L. V. Safris, "Calculation of magnetic amplifiers with respect to linearized magnetization characteristics," Tr. Rostovskogo-n/D Inst. Zh. -d. Transporta Vol. 20, Transzheldorizdat, 1956.

CHRONICLE

ALL-UNION CONFERENCE ON CONTACTLESS MAGNETIC ELEMENTS WHICH ARE USED IN AUTOMATION AND COMPUTER TECHNIQUES

G. V. Subbotina

Translated from Avtomatika i Telemekhanika, Vol. 22, No. 6,
pp. 811-814, June, 1961

An All-Union conference on contactless magnetic elements which are used in automation and computer techniques was held on February 20-23, 1961 in Minsk. This conference was organized by the Department of Solid State Physics and Semiconductors, Academy of Sciences, Belorus, SSR, the Scientific and Technological Division of the Belorussian Republic, The State Scientificotechnical Committee of the Council of Ministers, Belorus, SSR, and the Commission on Automation Contactless Magnetic Elements, Academy of Sciences, USSR. The aim of this conference was a discussion of problems in the theory, design, application, and development of contactless magnetic elements which are used in automation and computer techniques.

Over 400 representatives from scientific research organizations, higher institutes of learning, and industrial enterprises participated in this conference. Nearly 40 reports and communications, which aroused great interest among the participants at the conference, were read and discussed.

The president of the organizational committee of the conference, N. N. Sirota (Minsk), opened the conference

The reports by B. S. Sot-skov and M. A. Rozenblat were devoted to a survey of the contemporary state of the technology of contactless magnetic elements. In his report, entitled "The scope of electromechanical magnetic and semiconductor elements," B. S. Sot-skov (Moscow) discussed the main characteristics with respect to which elements of different types can be compared. These characteristics include the following: the input and output power, the actuation time, the input and output resistance, the multiplicity factor, the type of current, the number of circuits, the operating temperatures, the maximum allowable mechanical acceleration, reliability (the factor of breakdown danger), and the limits with respect to the number of operating cycles.

This comparison indicated the natural division of the fields where these elements are used. Besides the enumerated characteristics, among the other important characteristics are also power per unit volume, the lifetime before breakdown, the complexity of servicing, and the frequency with which a device breaks down. The technical and economic advantage in application depends on reliability, and finally electromechanical devices should be used in this order.

In his report "Modern contactless magnetic elements," M. A. Rozenblat (Moscow) proposed a possible classification of magnetic elements with respect to their operating principle: 1) Parametric elements, where the nonlinear character or the magnetization curve is utilized (ferromagnetic stabilizers, frequency multipliers and dividers, ferroresonance devices, etc.); 2) magnetic-modulation elements, i.e., elements with the action of two fields that differ with respect to their frequency (magnetic modulators, probes, magnetic amplifiers, magnetic relays, etc.); 3) hysteresis elements (memories, magnetic recording devices, counters, core integrators, etc.).

This report also indicated the main lines of the development of magnetic elements: increase in operating speed, improvement of efficiency and reduction of over-all dimensions and costs, improvement of technological efficiency, and dispensation with rectifiers and transistors in discrete-action magnetic elements.

The reports by N. N. Sirota and K. D. Mart'yanov were devoted to the present state of the development of magnetically soft materials and the physical processes occurring in them. In his report "Ferrites with a rectangular hysteresis loop and magnetic reversal processes in ferrites," N. N. Sirota presented the results obtained in investigating the dependence of ferrite properties on their composition. He also discussed the general laws governing the variation of characteristics in binary and ternary hard alloys. Besides, the composition, temperature, and the magnetizing field strength, the shape of the hysteresis loop also depends on structural uniformity, density, and the presence of cavities. Additional annealing broadens the loop, however, thermomagnetic treatment in the

region of the Curie point and the addition of small components make it possible to vary the characteristics of ferrites in the desired direction. The author briefly dealt with the mechanism of magnetization reversal processes in connection with the switching time and the composition of ferrites.

K. D. Mart'yanova (Moscow) reported on new developments in the field of soft ferromagnetic iron-nickel-cobalt alloys with a rectangular hysteresis loop and higher temperature stability (in the temperature range from 60 to 300° C).

The report by V. G. Baranovskii and I. A. Petrusenko (Leningrad) was devoted to the development of a push-pull magnetic bridge amplifier with positive feedback for even harmonics and with a common ac output. The operation of the proposed circuit, its basic properties, and the conditions for securing even harmonics in windings encompassing two cores were discussed. The results obtained in testing this circuit under steady-state and transient conditions when the temperature changed from -40 to +60°C were described.

This circuit can be controlled by direct as well as alternating current.

As an example of application, a servo system with a two-phase asynchronous motor, where a feedback circuit with even harmonics is used as the output stage, was discussed. The application of this circuit made it possible to reduce the size and weight of the output stage.

A number of reports were concerned with the results obtained in developing and investigating highly sensitive magnetic amplifiers and magnetic modulators. A. I. Moskalev (Moscow) reported on the development of a push-pull dc magnetic amplifier circuit with a single core, which made it possible to organize the industrial production of amplifiers with a zero stability of 1×10^{-13} w at 60°C.

In their report "A two-phase magnetic modulator with a low sensitivity threshold," F. I. Kerbnikov and M. A. Rozenblat (Moscow) described the application of the higher frequencies of magnetic-transistor oscillators in developing magnetic modulators with two-phase supply. The modulators were based on ferrite cores with mutually perpendicular excitation and signal fields. Two-phase feed modulators make it possible to dispense with the use of filters in the supply, control, and output circuits. For a supply frequency of 500 cps, an experimental model had a lower sensitivity threshold of 10^{-12} w, which was reduced to the 1-cps passband. M. A. Rakov and L. A. Sinitskii (L'vov) described similar modulators where two-phase supply is provided by means of two magnetic frequency dividers.

In their report "Magnetic pulse amplifiers," V. G. Anikin and I. A. Ikorokov (Moscow) dealt with the problem of improving the sensitivity of magnetic amplifiers with self-saturation by varying the supply voltage shape. Due to pulsed supply, the sensitivity can be increased to 10^{-10}–10^{-11} w. Methods for designing pulse sources with magnetic-transistor elements were also considered.

E. I. Gurvich and L. B. Shchukin (Moscow) described methods for measuring the characteristics of miniature memorizing transformers (with a diameter of 2-7 mm), which are to be used in contactless elements in automation and computer techniques.

The report by G. D. Kozlov (Moscow) was devoted to a method for calculating the basic core characteristics (the basic magnetization curve and the limiting and the particular magnetization reversal cycles) by approximating hysteresis loops with parallelograms.

The effect of the core shape on magnetization reversal processes was discussed in a communication presented by K. E. Volkovitskii (Moscow).

V. V. Sarv (Tallin) presented a report entitled "Acceleration of the growth of a magnetic set by means of accumulation and repeated utilization of the energy of a decreasing magnetic field."

E. L. L'vov (Moscow) discussed a method for determining the transfer function of choke-coupled magnetic amplifiers with resistive-inductive dc loads.

The author of this report obtained an expression for the transfer function of an idealized magnetic amplifier in terms of its circuit parameters and the input-putput characteristic. The expressions for the transfer and the transition functions were compared with the expressions obtained in other papers, and the parameter regions where the equations yield results different from experimental results were determined.

V. V. Kokhanov (Frunze) presented a communication on the development of a contactless automatic pulse generator which is designed for operation in combination with a data transmitter capable of converting nonelectrical quantities into ac voltage.

Many reports presented to the Conference were devoted to the results obtained in theoretical and experimental investigations and in the application of magnetic logical elements for industrial control systems.

N. P. Vasil'eva (Moscow) and I. Gaskovec (Prague) compared the operation of magnetic logical repeater circuits under steady-state conditions and under maximum-output conditions.

It was shown that a stable characteristic of circuits with decoupling by means of controllable diodes can be obtained by using a special supply voltage shape, while in circuits with decoupling by means of resistors, this can be achieved by a suitable choice of the circuit parameters.

It was also shown that, for a frequency of 50 cps, the decisive factors in determining the maximum number of elements are the conditions of the limiting magnetization reversal and nonreversal of magnetization in cores, while in the case of higher supply frequencies, the main factor is the heating of the control element core winding.

With respect to the above characteristics, the best circuit is that with decoupling by means of resistors.

V. S. Podlipenskii (Kiev) presented a communication entitled "Some theoretical relationships for logical-action magnetic elements." Elements based on the Reimi circuit were considered.

L. V. Shopen (Moscow) presented a calculation of magnetic logical elements (on the basis of push-pull resistors and a transformer circuit) for the case where the power is supplied by a sinusoidal voltage source.

I. A. Baranov and N. L. Prokhorov (Moscow) described the application of the magnetic logical elements that have been developed at the Institute of Automation and Remote Control, Academy of Sciences, USSR, in control circuits for controlling the electric and hydraulic drives of machine tool units and automatic assembly lines.

The report was devoted to methods for designing units and control circuits by using magnetic logical elements.

K. R. Lebedev (Kalinin) reported on logical element types whose mass-production is planned at the Kalinin plant for electrical equipment. The author described the supply sources and the control signal transmitters, and he provided information on the parameters of the elements as well their design.

B. V. Serov (Petrozavodsk) presented a report on the application of magnetic elements in devices for the automated sorting of heavy timber.

M. G. Kogan (Sverdlovsk) presented a report on control devices which are based on magnetic logical elements.

The report by V. S. Matorina and O. A. Sedykh (Moscow), entitled "Lag element for systems with magnetic logical elements," was devoted to various types of time lag elements, the operating principle of which is based on the dependence of the core magnetization reversal time on the voltage applied to the core winding.

The report also dealt with the possibility of using integrating cores for designing time lag elements.

M. M. Lotosh (Moscow) reported on the development of large smoothly controllable time lags which are based on magnetic cores.

Circuits which make it possible to obtain lag times of up to several hours were considered.

The author also provided a comparative analysis of circuits with low-frequency oscillators and of circuits with frequency dividers. Much attention was paid in the report to the development of a divider for dividing a number of pulses by a large factor.

A lag line consisting of a ferrite-triode cell was considered in the report by N. A. Yarmosh (Minsk).

The report by A. L. Pisareva and G. N. Derim-Oglu was devoted to the development of output magnetic amplifiers for contactless automation circuits. Contactors and ac electromagnets are most widely used as loads, which involves great difficulties in developing magnetic amplifiers due to sharp changes in the load in dependence on the magnetic system's position.

A series of magnetic amplifiers for the control of ac contactors with a power consumption of up to 350 va as well as for the control of ac apparatuses (TU series for 50 cps and TUA5 series for 400 cps) have been developed.

The report "Transformation of digital codes into analog codes by means of magnetic amplifiers," which was presented by M. A. Rozenblat and G. V. Subbotina (Moscow), was devoted to decision devices consisting of magnetic elements. This report was concerned with the theory and the results obtained in the development of a multiplace converter. The report by V. I. Semenova and É. F. Stepura (Leningrad), entitled "Application of magnetic amplifiers for the multiplication of two or several electrical signals" was devoted to the same subject.

In their report "Magnetic amplifier with combined feedback," V. G. Anikin and I. A. Okorokov (Moscow) described the development of an amplifier with large power gain and with high stability of the input-output response.

A. M. Bamdas, S. V. Shapiro, Yu. A. Savinovskii, and I. V. Blinov (Gor'kii reported on their work on logical frequency converters and on voltage stabilizers with magnetizable shunts, on the development of single-phase and three-phase voltage stabilizers with magnetizable autotransformers, and on new static frequency multipliers.

In their report "On the theory of parametrons," O. N. Sitnikov and G. Ya. Karasik (Sverdlovsk) analyzed the parametron operation by taking into account changes in the dynamic permeability of cores.

The authors demonstrated that the detuning of the parametrons is determined by the nonlinearity of the magnetic permeability and that a definite operating frequency range exists for a given material.

V. S. Toropov (Moscow) presented a report on a new method for information recording in magnetic memories.

A. B. Gorodetskii (Moscow) presented a report entitled "A promising direction in the development of magnetic amplifier technology," in which he offered a comparative survey of the existing designs and the parameters of domestic and foreign makes of magnetic amplifiers; he also set forth a number of proposals concerning the terminology and the methods of graphical representation of magnetic amplifier circuits.

V. M. Kurotchenko and N. I. Babanov (Frunze) reported on the development of the BKU-60 contactless code-pulse device for the telemetering of rotation angles.

Stefanovich, Shvarts, Gorodetskii, Isaev (Moscow), Atlasov, Petrusenko (Leningrad), Il'in, Chaikovskii (Minsk), Chizhukhin (Penza), Pashkovskii (L'vov), and others participated in the discussion which followed the reading of the reports. It should be mentioned that great interest was shown in the discussion concerning the problems in the technology and mass-production of standardized magnetic elements and units as well as the problems connected with perfecting the terminology.

At the conclusion of the conference, the participants adopted a resolution emphasizing the prospects offered by the use of contactless magnetic elements and the necessity of further scientific research and experimental work in this field. The resolution also proposed a number of measures to be taken with the aim of accelerating the introduction of the most advanced contactless elements into the national economy.

SIGNIFICANCE OF ABBREVIATIONS MOST FREQUENTLY ENCOUNTERED IN SOVIET TECHNICAL PERIODICALS

AN SSSR	*Academy of Sciences, USSR*
FIAN	*Physics Institute, Academy of Sciences USSR*
GITI	*State Scientific and Technical Press*
GITTL	*State Press for Technical and Theoretical Literature*
GOI	*State Optical Institute*
GONTI	*State United Scientific and Technical Press*
Gosénergoizdat	*State Power Press*
Gosfizkhimizdat	*State Physical Chemistry Press*
Goskhimizdat	*State Chemistry Press*
GOST	*All-Union State Standard*
Gostekhizdat	*State Technical Press*
GTTI	*State Technical and Theoretical Press*
IAT	*Institute of Automation and Remote Control*
IF KhI	*Institute of Physical Chemistry Research*
IFP	*Institute of Physical Problems*
IL	*Foreign Literature Press*
IPF	*Institute of Applied Physics*
IPM	*Institute of Applied Mathematics*
IREA	*Institute of Chemical Reagents*
ISN (Izd. Sov. Nauk)	*Soviet Science Press*
IYap	*Institute of Nuclear Studies*
Izd	*Press (publishing house)*
LÉTI	*Leningrad Electrotechnical Institute*
LFTI	*Leningrad Institute of Physics and Technology*
LIM	*Leningrad Institute of Metals*
LITMiO	*Leningrad Institute of Precision Instruments and Optics*
Mashgiz	*State Scientific-Technical Press for Machine Construction Literature*
MGU	*Moscow State University*
Metallurgizdat	*Metallurgy Press*
MOPI	*Moscow Regional Pedagogical Institute*
NIAFIZ	*Scientific Research Association for Physics*
NIFI	*Scientific Research Institute of Physics*
NIIMM	*Scientific Research Institute of Mathematics and Mechanics*
NIKFI	*Scientific Institute of Motion Picture Photography*
NKTM	*People's Commissariat of the Heavy Machinery Industry*
Obrongiz	*State Press of the Defense Industry*
OIYaI	*Joint Institute of Nuclear Studies*
ONTI	*United Scientific and Technical Press*
OTI	*Division of Technical Information*
OTN	*Division of Technical Science*
RIAN	*Radium Institute, Academy of Sciences of the USSR*
SPB	*All-Union Special Planning Office*
Stroiizdat	*Construction Press*
URALFTI	*Ural Institute of Physics and Technology*
TsNIITMASh	*Central Scientific Research Institute of Technology and Machinery*
VNIIM	*All-Union Scientific Research Institute of Metrology*

NOTE: Abbreviations not on this list and not explained in the translation have been transliterated, no further information about their significance being available to us – *Publisher.*

Publication of a "Soviet Instrumentation and Control Translation Series" by the Instrument Society of America has been made possible by a grant in aid from the National Science Foundation, with additional assistance from the National Bureau of Standards for the journal *Measurement Techniques.*

Subscription rates have been set at modest levels to permit widest possible distribution of these translated journals.

The Series now includes four important Soviet instrumentation and control journals. The journals included in the Series, and the subscription rates for the translations, are as follows:

MEASUREMENT TECHNIQUES – *Izmeritel'naya Tekhnika*

Russian original published by the Committee of Standards, Measures, and Measuring Instruments of the Council of Ministers, USSR. The articles in this journal are of interest to all who are engaged in the study and application of fundamental measurements. Both 1958 (bimonthly) and 1959-1961 (monthlies) available.

Per year (12 issues) starting with 1961, No. 1

General:	United States and Canada . .	$25.00
	Elsewhere	28.00
Libraries of nonprofit academic institutions:		
	United States and Canada . .	$12.50
	Elsewhere	15.50

INSTRUMENTS AND EXPERIMENTAL TECHNIQUES
Pribory i Tekhnika Éksperimenta

Russian original published by the Academy of Sciences, USSR. The articles in this journal relate to the function, construction, application, and operation of instruments in various fields of experimentation. 1958-1961 issues available.

Per year (6 issues) starting with 1961, No. 1

General:	United States and Canada . .	$25.00
	Elsewhere	28.00
Libraries of nonprofit academic institutions:		
	United States and Canada . .	$12.50
	Elsewhere	15.50

AUTOMATION AND REMOTE CONTROL – *Avtomatika i Telemekhanika*

Russian original published by the Institute of Automation and Remote Control of the Academy of Sciences, USSR. The articles are concerned with analysis of all phases of automatic control theory and techniques. 1957-1961 issues available.

Per year (12 issues) starting with Vol. 22, No. 1

General:	United States and Canada . .	$35.00
	Elsewhere	38.00
Libraries of nonprofit academic institutions:		
	United States and Canada . .	$17.50
	Elsewhere	20.50

INDUSTRIAL LABORATORY – *Zavodskaya Laboratoriya*

Russian original published by the Ministry of Light Metals, USSR. The articles in this journal relate to instrumentation for analytical chemistry and to physical and mechanical methods of materials research and testing. 1958-1961 issues available.

Per year (12 issues) starting with Vol. 27, No. 1

General:	United States and Canada . .	$35.00
	Elsewhere	38.00
Libraries of nonprofit academic institutions:		
	United States and Canada . .	$17.50
	Elsewhere	20.50

Single issues of all four journals, to everyone, each . . . $6.00

Prices on 1957-1960 issues available upon request

SPECIAL SUBSCRIPTION OFFER:

One year's subscription to all four journals of the 1961 Series, as above listed:

General:	United States and Canada . .	$110.00
	Elsewhere	122.00
Libraries of nonprofit academic institutions:		
	United States and Canada . .	$ 55.00
	Elsewhere	67.00

Subscriptions should be addressed to:

Instrument Society of America
530 William Penn Place
Pittsburgh 19, Penna.

1925